BIOINFORMATICS

Vol. 24 ISMB 2008, pages i232–i240
doi:10.1093/bioinformatics/btn162

Prediction of drug–target interaction networks from the integration of chemical and genomic spaces

Yoshihiro Yamanishi [1,*,†], Michihiro Araki [2], Alex Gutteridge [1], Wataru Honda [1] and Minoru Kanehisa [1,2]

[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011 and [2]Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai Minato-ku, Tokyo 108-8639, Japan

ABSTRACT

Motivation: The identification of interactions between drugs and target proteins is a key area in genomic drug discovery. Therefore, there is a strong incentive to develop new methods capable of detecting these potential drug–target interactions efficiently.

Results: In this article, we characterize four classes of drug–target interaction networks in humans involving enzymes, ion channels, G-protein-coupled receptors (GPCRs) and nuclear receptors, and reveal significant correlations between drug structure similarity, target sequence similarity and the drug–target interaction network topology. We then develop new statistical methods to predict unknown drug–target interaction networks from chemical structure and genomic sequence information simultaneously on a large scale. The originality of the proposed method lies in the formalization of the drug–target interaction inference as a supervised learning problem for a bipartite graph, the lack of need for 3D structure information of the target proteins, and in the integration of chemical and genomic spaces into a unified space that we call 'pharmacological space'. In the results, we demonstrate the usefulness of our proposed method for the prediction of the four classes of drug–target interaction networks. 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.

Availability: Softwares are available upon request.

Contact: Yoshihiro.Yamanishi@ensmp.fr

Supplementary information: Datasets and all prediction results are available at http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/.

1 INTRODUCTION

The identification of 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. Through various high-throughput experimental projects for analyzing the genome, transcriptome and proteome, we are beginning to understand the genomic spaces populated by these protein classes. At the same time, the high-throughput screening of large-scale chemical compound libraries with various biological assays is enabling us to explore the chemical space of possible compounds (Dobson, 2004; Kanehisa *et al.*, 2006; Stockwell, 2000). The aim of chemical genomics research is to relate this chemical space with the genomic space in order to identify potentially useful compounds such as imaging probes and drug leads. However, our knowledge about the relationship between the chemical and genomic spaces is very limited. The PubChem database at NCBI (Wheeler *et al.*, 2006), for example, stores information on millions of chemical compounds, but the number of compounds with information on their target protein is very limited. This implies that many potential interactions between the chemical and genomic spaces remain undiscovered. Therefore, there is a strong incentive to develop new methods capable of detecting these potential drug–target interactions efficiently.

Since experimental determination of compound–protein interactions or potential drug–target interactions remains very challenging (Haggarty *et al.*, 2003; Kuruvilla *et al.*, 2002), effective *in silico* prediction methods need to be developed. The predicted interactions can provide complementary and supporting evidence to experimental studies. A variety of computational approaches have been developed to analyze and predict compound–protein interactions. Two of the most commonly used are docking simulations (Cheng *et al.*, 2007; Rarey *et al.*, 1996) and literature text mining (Zhu *et al.*, 2005). However, both techniques have their limitations, docking, for instance, cannot be applied to proteins whose 3D structures are unknown, so it is difficult to use this technique on a large scale. Text mining approaches are usually based on keyword searching and so suffer from an inability to detect new biological findings and also the problem of redundancy in the compound/gene names in the literature (Zhu *et al.*, 2005).

Recently, a classification of target proteins based on the structure of their ligands (Keiser *et al.*, 2007) and in related work an analysis of the drug–target network revealed characteristic features of its network topology (Yildirim *et al.*, 2007). However, neither protein sequence information nor chemical structure information were taken into consideration in the network analysis. The next step is to develop more integrative methods taking into account target protein sequences, drug chemical structures and the available known drug–target network information simultaneously.

In this article, we investigate the relationship between drug chemical structure, target protein sequence and drug–target network topology. We then develop a new supervised method to infer unknown drug–target interactions by integrating chemical space and genomic space into a unified space that we call 'pharmacological space'. In the proposed method, chemical space means the chemical structure similarity space of possible chemical compounds, genomic space means the amino acid sequence similarity space of possible

*To whom correspondence should be addressed.

†Present address: Centre for Computational Biology, Ecole des Mines de Paris, 35 rue Saint Honore, 77305 Fontainebleau Cedex, France.

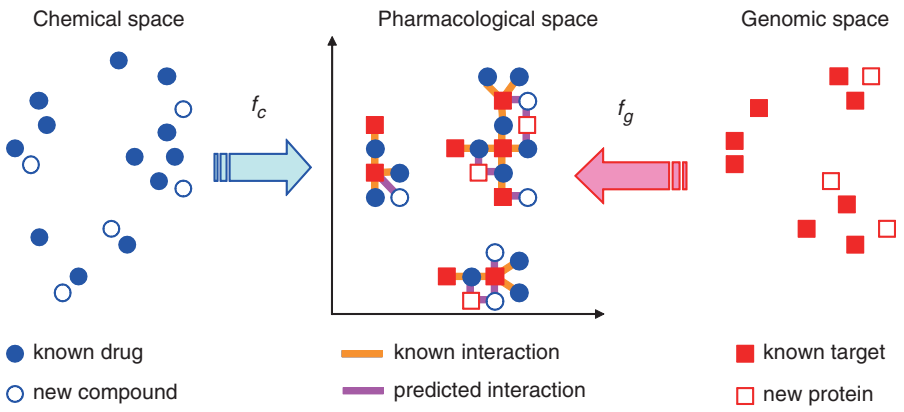

Fig. 1. An illustration of the proposed method.

proteins and pharmacological space means the interaction space reflecting the drug–target interaction network, where interacting drugs and target proteins are close to each other. By supervised we mean that reliable a priori knowledge about known interactions is used in the inference process itself. Figure 1 shows an illustration of our method. To our knowledge, there are no computational methods to predict drug–target interactions from the integration of chemical structure data, genomic sequence data and known drug–target network information simultaneously on a large scale. In the results, we make predictions for four classes of important drug–target interactions in human 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

2.1 Drug–target interaction data

We obtained the information about the interactions between drugs and target proteins from the KEGG BRITE (Kanehisa *et al.*, 2006), BRENDA (Schomburg *et al.*, 2004), SuperTarget (Gunther *et al.*, 2008) and DrugBank databases (Wishart *et al.*, 2008). According to our survey, the number of known drugs targeting enzymes, ion channels, GPCRs and nuclear receptors are 445, 210, 223 and 54, respectively. At the time of writing, the number of target proteins in these classes are 664, 204, 95 and 26, respectively, and the number of known interactions are 2926, 1476, 635 and 90. Note that in the enzyme class we focused on the regulatory interactions between enzymes and compounds rather than the metabolic interactions, so all the ligands in the enzyme data are inhibitors or activators rather than substrates or products. Cofactors such as adenosine triphosphate (ATP) and nicotinamide adenine dinucleotide phosphate (NADPH) are also not included except when they are annotated as regulators in the BRENDA database. Also, we do not use compounds whose molecular weights are <100, which means that ions are removed from the dataset. The data statistics for drugs and target proteins and their interactions are summarized in Table 1.

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.

Table 1. Statistics for the drug–target interaction networks

Statistics	Enzyme	Ion channel	GPCR	Nuclear receptor
No. of drugs	445	210	223	54
No. of target proteins	664	204	95	26
(Total in human genome)	(2741)	(292)	(757)	(49)
No. of drug–target interactions	2926	1476	635	90
Average degree of drugs	6.57	7.02	2.84	1.66
Average degree of targets	4.40	7.23	6.68	3.46
Cluster coefficient of drugs	0.850	0.871	0.867	0.832
Cluster coefficient of targets	0.902	0.897	0.776	0.933
Proportion of unreachable paths between drugs	0.479	0.019	0.345	0.615
Proportion of unreachable paths between targets	0.447	0.029	0.593	0.778

Table 1 shows the number of target proteins, drugs and their interactions in the gold standard data.

2.2 Chemical data

Chemical structures of the drugs were obtained from the DRUG and COMPOUND Sections in the KEGG LIGAND database (Kanehisa *et al.*, 2006). We computed the chemical structure similarities between compounds using SIMCOMP (Hattori *et al.*, 2003), where SIMCOMP provides a global similarity score based on the size of the common substructures between two compounds using a graph alignment algorithm. The similarity between two compounds c and c' is computed as $s_c(c,c')=|c\cap c'|/|c\cup c'|$. Applying this operation to all compound pairs, we construct a similarity matrix denoted as S_c. The similarity matrix S_c is considered to represent chemical space.

2.3 Genomic data

Amino acid sequences of the target proteins were obtained from the KEGG GENES database (Kanehisa *et al.*, 2006). In this study we focused on the proteins in human. We computed the sequence similarities between the proteins

using a normalized version of Smith–Waterman scores (Smith and Waterman, 1981). The normalized Smith–Waterman score between two proteins g and g' is computed as $s_g(g,g')=SW(g,g')/\sqrt{SW(g,g)}\sqrt{SW(g',g')}$, where $SW(\cdot,\cdot)$ means the original Smith–Waterman score. Applying this operation to all protein pairs, we construct a similarity matrix denoted as S_g. In this study the similarity matrix S_g is considered to represent genomic space.

3 METHODS

The proposed supervised method is a two-step process. First, a model is learned to explain the 'gold standard'. Second, this model is applied to compounds and proteins absent from the 'gold standard' in order to infer their interactions. A supervised learning method is suitable in this case, because information about reliable drug–target interactions is available from many public databases recently. The set of compounds and proteins involved in the known drug–target interactions are referred to as the training set. We first propose two 'naive' approaches: the nearest profile method and the weighted profile method, and we finally propose a more sophisticated approach: the bipartite graph learning method.

In each case, suppose that we have sets of known drugs $\{c_i\}_{i=1}^{n_c}$ and known target proteins $\{g_j\}_{j=1}^{n_g}$, where n_c is the number of known drugs and n_g is the number of known target proteins. Also, the interaction patterns of c_i with target proteins and g_j with drugs are represented by bit strings that we call the interaction profiles $\mathbf{x}_{c_i}$ and $\mathbf{y}_{g_j}$, respectively. The interaction profile $\mathbf{x}_{c_i}$ is defined as a bit string (vector of size n_g), where the presence or absence of an interaction with target protein $g_j (j=1,2,\ldots,n_g)$ is coded as 1 or 0, respectively. The interaction profile $\mathbf{y}_{g_j}$ is defined as a bit string (vector of size n_c), where the presence or absence of an interaction with drug $c_i (i=1,2,\ldots,n_c)$ is coded as 1 or 0, respectively. Suppose that we have sets of interaction profiles $\{\mathbf{x}_{c_i}\}_{i=1}^{n_c}$ and $\{\mathbf{y}_{g_j}\}_{j=1}^{n_g}$. Given a new target candidate protein g_{new} and a new drug candidate compound c_{new}, we want to predict the corresponding interaction profiles $\mathbf{x}_{c_{\mathrm{new}}}$ and $\mathbf{y}_{g_{\mathrm{new}}}$, respectively.

3.1 Nearest profile method

A straightforward approach is to use the idea of the nearest neighbor method. In this method, we predict the new compound c_{new} to have the following interaction profile:

$$\mathbf{x}_{c_{\mathrm{new}}} = s_c(c_{\mathrm{new}}, c_{\mathrm{nearest}})\mathbf{x}_{c_{\mathrm{nearest}}}, \tag{1}$$

where $\mathbf{x}_c$ is an interaction profile vector, $s_c(\cdot,\cdot)$ is a chemical similarity score, and c_{nearest} is the nearest compound which is most similar to c_{new}. We predict the new protein g_{new} to have the following interaction profile:

$$\mathbf{y}_{g_{\mathrm{new}}} = s_g(g_{\mathrm{new}}, g_{\mathrm{nearest}})\mathbf{y}_{g_{\mathrm{nearest}}}, \tag{2}$$

where $\mathbf{y}_g$ is an interaction profile vector, $s_g(\cdot,\cdot)$ is a sequence similarity score and g_{nearest} is the nearest protein which is most similar to g_{new}. Finally, high scoring compound–protein pairs (c_{new}, g_j) and (c_i, g_{new}) in the interaction profiles $\mathbf{x}_{c_{\mathrm{new}}}$ and $\mathbf{y}_{g_{\mathrm{new}}}$ are predicted to interact with each other. The method is referred to as nearest profile method in this study.

3.2 Weighted profile method

We consider a more generalized version of the above method. In this method, we predict the new compound c_{new} to have the following weighted interaction profile:

$$\mathbf{x}_{c_{\mathrm{new}}} = \frac{1}{z_{c_{\mathrm{new}}}} \sum_{i=1}^{n_c} s_c(c_{\mathrm{new}}, c_i)\mathbf{x}_{c_i}, \tag{3}$$

where $\mathbf{x}_c$ is an interaction profile vector, $s_c(\cdot,\cdot)$ is a chemical structure similarity score and $z_{c_{\mathrm{new}}}$ is a normalization term defined as $z_{c_{\mathrm{new}}} = \sum_{i=1}^{n_c} s_c(c_{\mathrm{new}}, c_i)$. We predict the new protein g_{new} to have the folowing weighted interaction profile:

$$\mathbf{y}_{g_{\mathrm{new}}} = \frac{1}{z_{g_{\mathrm{new}}}} \sum_{j=1}^{n_g} s_g(g_{\mathrm{new}}, g_j)\mathbf{y}_{g_j}, \tag{4}$$

where $\mathbf{y}_g$ is an interaction profile vector, $s_g(\cdot,\cdot)$ is a sequence similarity score and $z_{g_{\mathrm{new}}}$ is a normalization term defined as $z_{g_{\mathrm{new}}} = \sum_{j=1}^{n_g} s_g(g_{\mathrm{new}}, g_j)$. Finally, high-scoring compound–protein pairs (c_{new}, g_j) and (c_i, g_{new}) in the interaction profiles $\mathbf{x}_{c_{\mathrm{new}}}$ and $\mathbf{y}_{g_{\mathrm{new}}}$ are predicted to interact with each other. The method is referred to as weighted profile method in this study.

3.3 Bipartite graph learning method

The novel method used in this article is the bipartite graph learning method. Here we propose a new method to learn the correlation between the chemical/genomic space and the interaction space that we call 'pharmacological space'. The proposed procedure is as follows:

(1) Embed compounds and proteins on the interaction network into a unified space that we call 'pharmacological space'.

(2) Learn a model between the chemical/genomic space and the pharmacological space, and map any compounds/proteins onto the pharmacological space.

(3) Predict interacting compound–protein pairs by connecting compounds and proteins which are closer than a threshold in the pharmacological space.

Figure 1 shows an illustration of the above procedure. The details of each step are explained below.

First, the drug–target interaction network is described by a bipartite graph $G=(V_1+V_2,E)$, where V_1 is a set of drugs, V_2 is a set of target proteins and E is a set of the interactions. We propose to represent the bipartite graph structure by an Euclidian space such that both compounds and proteins are represented by sets of q-dimensional feature vectors $\{\mathbf{u}_{c_i}\}_{i=1}^{n_c}$ and $\{\mathbf{u}_{g_j}\}_{j=1}^{n_g}$, respectively. To do so, we first construct a graph-based similarity matrix $K=\begin{pmatrix} K_{cc} & K_{cg} \\ K_{cg}^T & K_{gg} \end{pmatrix}$, where the elements of K_{cc}, K_{gg} and K_{cg} are computed by using Gaussian functions as follows: $(K_{cc})_{ij}=\exp(-d^2_{c_ic_j}/h^2)$ for $i,j=1,\ldots,n_c$, $(K_{gg})_{ij}=\exp(-d^2_{g_ig_j}/h^2)$ for $i,j=1,\ldots,n_g$ and $(K_{cg})_{ij}=\exp(-d^2_{c_ig_j}/h^2)$ for $i=1,\ldots,n_c, j=1,\ldots,n_g$, where d is the shortest distance between all objects (compounds and proteins) on the bipartite graph, the distance between unreachable object pairs is treated as infinity and h is a width parameter. Note that the size of the resulting matrix K is $(n_c+n_g)\times(n_c+n_g)$. The matrix K is not always positive definite, so an appropriate identity matrix is added to the K such that the matrix K meets the positive definite property. Borrowing a similar idea with kernel principal component analysis (Scholkopf *et al.*, 1998), we apply the eigenvalue decomposition of K as $K=\Gamma\Lambda^{1/2}\Lambda^{1/2}\Gamma^T=UU^T$, where the diagonal elements of matrix Λ are eigenvalues and columns of matrix Γ are eigenvectors and $U=\Gamma\Lambda^{1/2}$. Then, we represent all drugs and target proteins by using the row vectors of the matrix $U=(\mathbf{u}_{c_1},\ldots,\mathbf{u}_{c_{n_c}},\mathbf{u}_{g_1},\ldots,\mathbf{u}_{g_{n_g}})^T$. The space spanned by features $\mathbf{u}_c$ and $\mathbf{u}_g$ is referred to as 'pharmacological feature space'.

Second, we consider a model representing the correlation between the chemical/genomic space and the pharmacological feature space. To do so, we propose to apply a variant of the kernel regression model $f:\mathcal{X}\times\mathcal{X}\to\mathbf{R}^q$ as follows:

$$\mathbf{u}=f(x,x_i)=\sum_{i=1}^{n} s(x,x_i)\mathbf{w}_i+\epsilon, \tag{5}$$

where x is an object belonging to a set $\mathcal{X}$, n is the size of the set $\mathcal{X}$, f is the projection from a similarity space to a Euclidean space, $s(\cdot,\cdot)$ is a similarity score function, $\mathbf{w}_i$ is a weight vector and ϵ is a noise vector. The optimization can be done by finding $\mathbf{w}_i$ which minimizes the following loss function:

$$L=||UU^T-SWW^TS^T||_F^2, \tag{6}$$

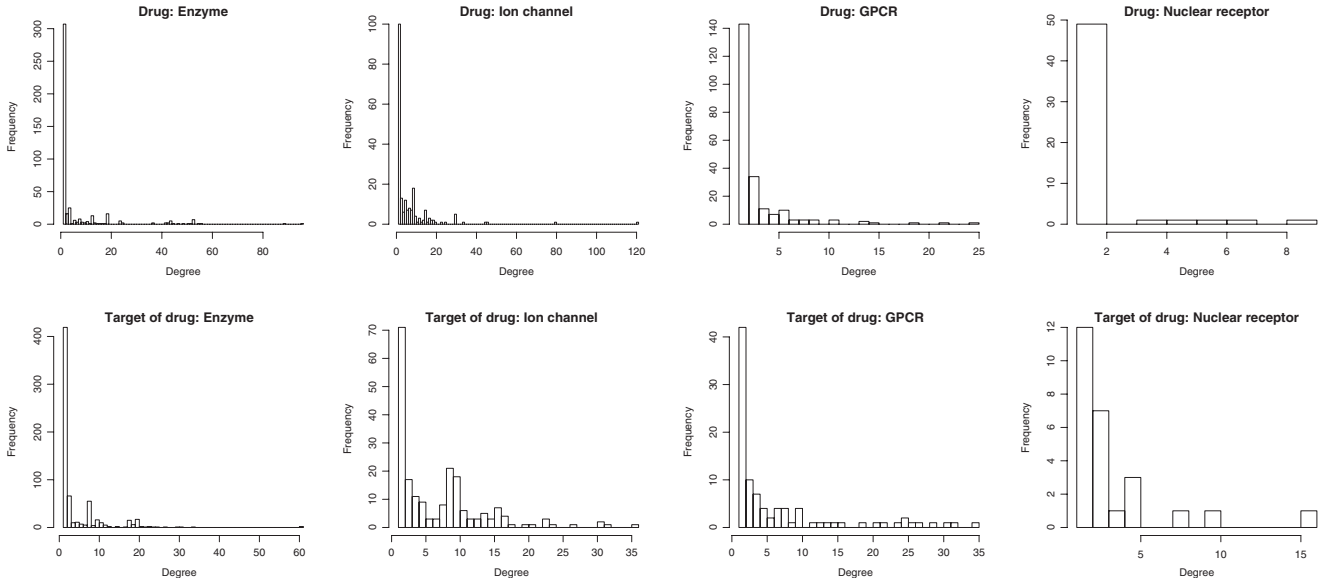

Fig. 2. Degree distributions for drugs and target proteins. The top four panels show the histograms of the degree of drugs targeting enzyme, ion channel, GPCR and nuclear receptor, respectively. The bottom four panels show the histogram of the degree of the corresponding target proteins.

where S is an $n \times n$ similarity matrix, $W = (\mathbf{w}_1, \ldots, \mathbf{w}_n)^T$, and $||\cdot||_F$ is Frobenius norm. In this study, we learn two models: f_c for the correlation between the chemical space and the pharmacological feature space and f_g for the correlation between the genomic space and the pharmacological feature space, respectively. Suppose that we have a new compound c_{new} and a new protein g_{new}. Applying the model f_c, we map the new compound c_{new} onto the pharmacological feature space as

$$\mathbf{u}_{c_{\text{new}}} = f_c(c_{\text{new}}, c_i) = \sum_{i=1}^{n_c} s_c(c_{\text{new}}, c_i)\mathbf{w}_{c_i}, \tag{7}$$

where $\mathbf{w}_{c_i}$ is a weight vector and $s_c(\cdot,\cdot)$ is a chemical structure similarity score. Applying the model f_g, we map the new protein g_{new} onto the pharmacological feature space as

$$\mathbf{u}_{g_{\text{new}}} = f_g(g_{\text{new}}, g_j) = \sum_{j=1}^{n_g} s_g(g_{\text{new}}, g_j)\mathbf{w}_{g_j}, \tag{8}$$

where $\mathbf{w}_{g_j}$ is a weight vector and $s_g(\cdot,\cdot)$ is a sequence similarity score.

Finally, based on the features in the pharmacological space, we compute the feature-based similarity scores for three types of compound–protein pairs by calculating the inner product as follows: (i) $corr(c_{\text{new}}, g_j) = \mathbf{u}_{c_{\text{new}}} \cdot \mathbf{u}_{g_j}$, (ii) $corr(c_i, g_{\text{new}}) = \mathbf{u}_{c_i} \cdot \mathbf{u}_{g_{\text{new}}}$ and (iii) $corr(c_{\text{new}}, g_{\text{new}}) = \mathbf{u}_{c_{\text{new}}} \cdot \mathbf{u}_{g_{\text{new}}}$. The feature-based similarity score is used as a measure of the closeness between compounds and proteins in the pharmacological feature space. Then, high-scoring compound–protein pairs are predicted to interact with each other.

4 RESULTS

4.1 Drug–target interaction network construction

In this study we focus on interactions made by four pharmaceutically useful drug–target classes: enzymes, ion channels, GPCRs and nuclear receptors. We constructed the drug–target interaction network for each protein class using a bipartite graph representation. 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 degree distributions for drugs and target proteins in the drug–target interaction network. The degree of the drug (respective protein) node is the number of targets that the drug has (respectively the number of drugs targeting the protein). Among the four classes, ion channels and their corresponding drugs have many nodes with large degree, compared with the other protein classes.

Table 1 also shows the average degree, the clustering coefficient, and the proportion of unreachable paths for the drug–drug, target–target and drug–target pairs. The high values of the clustering coefficients imply that drugs and their targets tend to be densely clustered in the drug–target networks. We observe that the proportion of unreachable paths in the ion channel network tends to be smaller than those in the other protein classes, implying that most compound–protein pairs are connected in the network. Inspection of the network shows that the enzyme, GPCR and nuclear receptor networks comprise many small unconnected components, while the ion channel network tends to form one giant connected component. This also suggests that enzymes, GPCRs and nuclear receptor have strong binding specificity with their ligands, compared with ion channels.

4.2 Relation with chemical space and genomic space

We also investigated how the network topology is related to the chemical and genomic spaces. We used the SIMCOMP score to measure the chemical structure similarity between compounds, and we used the normalized Smith–Waterman score to measure the sequence similarity between target proteins.

Figure 3 shows the distributions of drug–drug chemical structural similarities and target–target sequence similarities against their distances in the drug–target interaction network for the four classes of targets. From the figure we observe several features. First, the larger the network distance between drugs and between targets, the smaller the variability of drug structure similarities and target sequence similarities, respectively. Second, the larger the network distance, the lower the averages of the drug structure similarity and the target sequence similarity. These observations imply that two

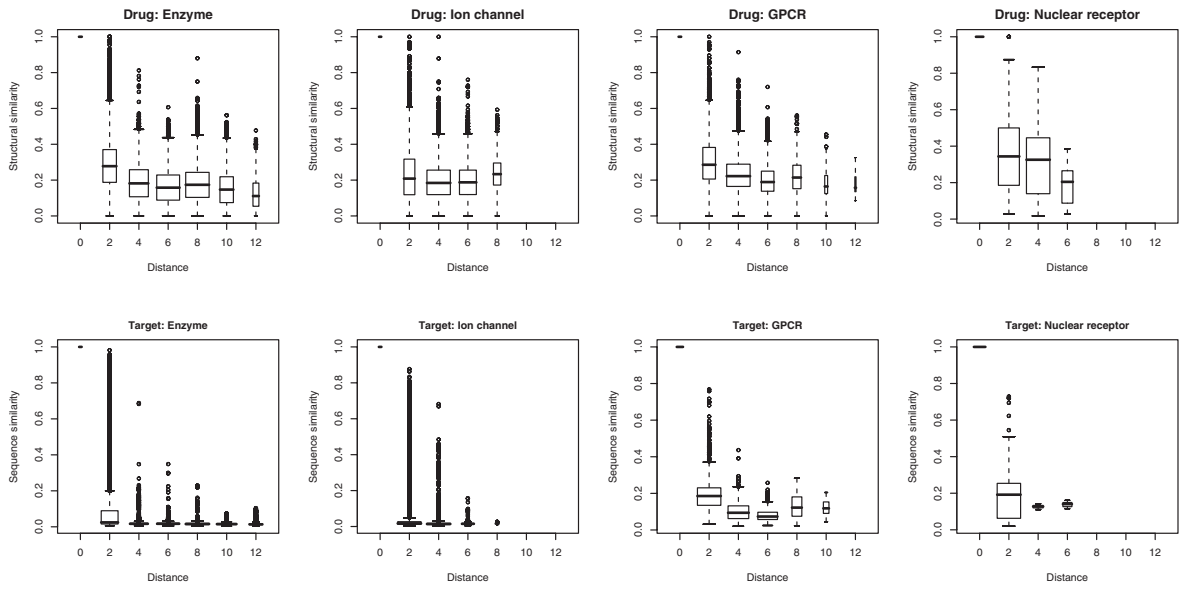

Fig. 3. Box-plots of chemical structure similarities between drugs and sequence similarities between target proteins against the network distance for enzyme, ion channel, GPCR, and nuclear receptor, respectively. The top four panels show the box-plot of the SIMCOMP scores between drugs against the network distance (d = 0, 2, 4, 6, …). The bottom four panels show the box-plot of the normalized Smith–Waterman scores between target proteins against the network distance (d = 0, 2, 4, 6, …). Note that the distance means the shortest path between objects (drugs or target proteins in each case) on the bipartite graph representation for the drug–target interaction network.

compounds sharing high structure similarity tend to interact with similar target proteins. Likewise two target proteins sharing high sequence similarity tend to interact with similar drugs and hence are close in the network. These observations suggest a strong correlation between interaction partners, structural similarities of drugs and the sequence similarities of target proteins.

4.3 Performance evaluation of the proposed methods

The three methods: 'nearest profile', 'weighted profile' and 'bipartite graph learning' were tested on the four classes of drug–target interactions involving enzymes, ion channels, GPCRs and nuclear receptors. We performed the following 10-fold cross-validation procedure: the gold standard set was split into 10 subsets of roughly equal size, each subset was then taken in turn as a test set, and we performed the training on the remaining nine sets. The performance was evaluated by using a receiver operating curve (ROC; Gribskov and Robinson, 1996), that 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. In the bipartite graph learning method we set parameter h to 2 in each protein class, because the cross-validation experiment provided the best prediction accuracy with $h = 2$.

Figure 4 shows the ROC curves of the bipartite graph learning method for the four classes of drug–target interactions. For each drug–target interaction class, the ROC curves are drawn for different sets of predictions depending on whether the compound and/or the protein were in the initial training set or not. Compounds and proteins in the training set are called 'known' whereas those not in the training set are called 'new'. Four different classes are then possible: (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) all the possible predictions (the average of the above three parts), which are colored red, green, blue and black, respectively. The bipartite graph learning method seems to catch sufficient information to detect all four types of drug–target interactions at high true-positive rates against low false-positive rates at any threshold. Among the four classes of drug–target interactions, the proposed method seems to have highest prediction ability for enzymes and GPCR, followed by ion channels and nuclear receptors. As one would expect, predictions where neither the protein nor the compound are in the training set (iii) are weakest, but even then reliable predictions are possible.

We compared the performance between the methods using several statistics. Table 2 shows the AUC (area under the ROC curve), sensitivity, specificity and PPV (positive predictive value) when the upper one percentile in the prediction score is chosen as a threshold, because high-confidence prediction results are interesting in practical applications. All the methods have quite high specificity, but the other statistics vary. The bipartite graph learning method outperforms the other methods with not only high AUC, but also high sensitivity and PPV. One explanation for the low sensitivity of the nearest profile and weighted profile methods is that they cannot predict interactions between new drug candidate compounds and new target candidate proteins [prediction class (iii) earlier], while this is possible with the bipartite graph learning method. These results serve to highlight the significant performance of the bipartite graph learning method.

4.4 Comprehensive prediction for unknown drug–target interactions

After confirming the usefulness of our method we conducted a comprehensive prediction of interactions between all possible

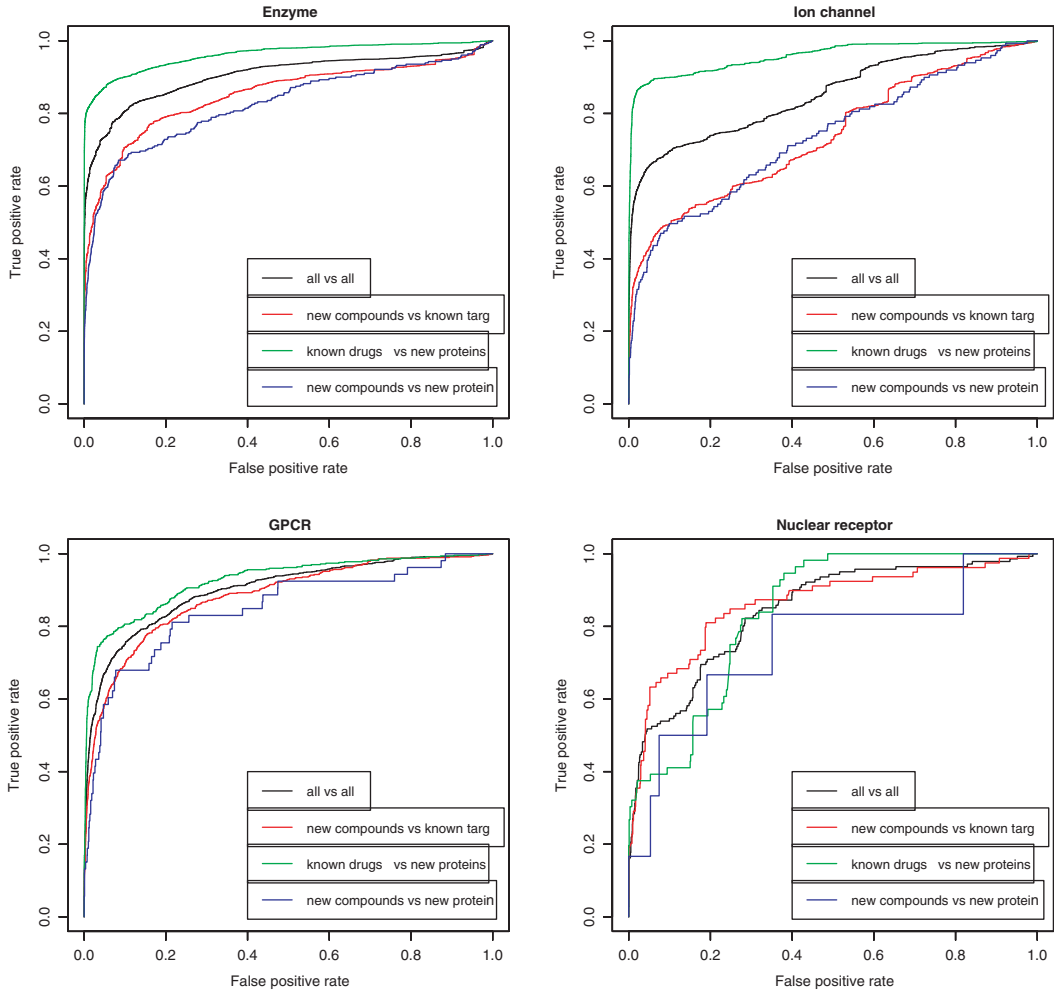

Fig. 4. ROC curves of the bipartite graph learning method for four classes of drug–target interactions: enzymes, ion channels, GPCRs and nuclear receptors.

Table 2. Statistics of the prediction performance

Data	Method	AUC	Sensitivity	Specificity	PPV
Enzyme	Nearest profile	0.767	0.538	0.995	0.532
	Weighted profile	0.812	0.386	0.993	0.384
	Bipartite graph learning	0.904	0.574	0.995	0.570
Ion channel	Nearest profile	0.751	0.166	0.995	0.576
	Weighted profile	0.811	0.239	0.998	0.826
	Bipartite graph learning	0.851	0.271	0.999	0.936
GPCR	Nearest profile	0.729	0.156	0.994	0.474
	Weighted profile	0.739	0.146	0.994	0.444
	Bipartite graph learning	0.899	0.234	0.996	0.681
Nuclear receptor	Nearest profile	0.710	0.073	0.993	0.440
	Weighted profile	0.626	0.114	0.998	0.818
	Bipartite graph learning	0.843	0.148	0.999	0.954

The AUC (ROC score) is the area under the ROC curve, 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 is defined as TP/(TP+FP), here TP, FP, TN, FN are the number of true positives, false positives, true negatives and false negatives, respectively.

compounds and proteins for the four classes of target proteins studied: enzymes, ion channels, GPCRs and nuclear receptors. 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 between all human proteins annotated as members of the four classes in KEGG GENES and all compounds in KEGG LIGAND. According to our survey based on the KEGG database, the number of enzymes, ion channels, GPCR nuclear receptors coded in the human genome are at least 2741, 292, 757 and 49, respectively, while the number of compounds used for the prediction is 15383 in each case. All the prediction results and high resolution graph pictures can be obtained from the web supplement. Because of space limitations, we have focused on the results for enzymes and GPCRs below.

4.4.1 Predicted enzymes interaction network Figure 5 shows a partial graph of the predicted network for enzyme data, where the top 100 scoring predictions are shown. Table 3 shows some examples of predicted enzyme-compound pairs with high interaction scores. The top scoring predictions for the enzyme dataset are dominated by interactions involving a few enzyme and compound families. These families tend to be those where the enzymes

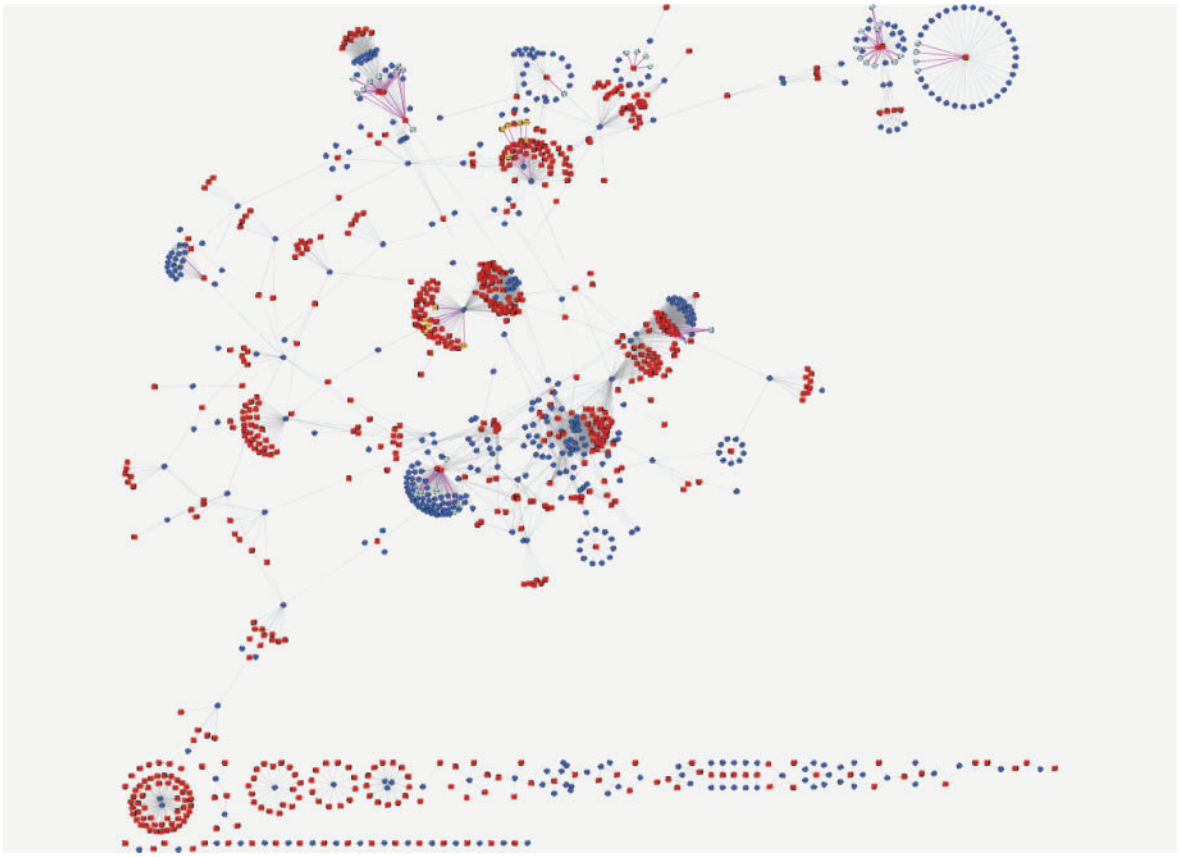

Fig. 5. Predicted enzymes interaction network. Blue, red, light blue and orange nodes indicate known drugs, known targets, newly predicted compounds and newly predicted proteins, respectively. Gray and pink edges indicate known interactions and newly predicted interactions with 100 highest scores, respectively.

Table 3. Top scoring predicted compound–protein pairs for enzyme data

Rank	Score	Pair	Annotation
1	0.924	C06977	Enalapril
		1636	angiotensin I converting enzyme 1
2	0.857	D01441	Imatinib mesilate (JAN)
		2444	fyn-related kinase [EC:2.7.10.2]
3	0.857	D00160	Epsilon-Aminocaproic acid (JAN)
		5644	protease, serine, 1 (trypsin 1) [EC:3.4.21.4]
4	0.844	C11720	Enalaprilate
		1636	angiotensin I converting enzyme 1
5	0.833	D00160	Epsilon-Aminocaproic acid (JAN)
		7177	tryptase alpha/beta 1 [EC:3.4.21.59]
6	0.824	D00043	Isoflurophate (USP)
		5644	protease, serine, 1 (trypsin 1) [EC:3.4.21.4]
7	0.81	D01605	Meticrane (JP15)
		759	carbonic anhydrase I [EC:4.2.1.1]
8	0.81	D00043	Isoflurophate (USP)
		7177	tryptase alpha/beta 1 [EC:3.4.21.59]
9	0.809	D00160	Epsilon-Aminocaproic acid (JAN)
		440387	chymotrypsinogen B2 [EC:3.4.21.1]
10	0.807	D01441	Imatinib mesilate (JAN)
		5753	PTK6 protein tyrosine kinase 6 [EC:2.7.10.2]

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

are both druggable and widely studied, or were a single initial drug compound has been developed into many derivatives, leading to a wealth of compound binding information being available for them. The six commonest enzyme families are angiotensin converting enzyme (ACE), tyrosine kinases, trypsin-related serine proteases, carbonic anhydrases, cyclooxygenases (COX) 1/2 and topoisomerases. Interactions with these six families account for 49 out of the top 50 predictions. Some of the predictions are trivial, particularly where many chemically almost identical compounds are available in the dataset, but interesting cases also come up.

COX enzymes are a common target for antiinflammatory drugs due to their role in the synthesis of prostanoids and the subsequent inflammation response (Rainsford, 2007). Amongst the top predictions for COX is a known antiinflammatory drug Cicloprofen (D03489), so the high predicted score is encouraging. Two potentially novel COX interactions are also predicted with 4-hydroxyhydratropate (C03080) and 2,2-*bis*(4-hydroxyphenyl)-propanoic acid (C13633), neither of which have previously been identified as potential COX inhibitors to our knowledge.

A compound that appears several times in the top 50 predictions is Imatinib mesilate (D01441), a tyrosine kinase inhibitor used in the treatment of chronic myelogenous leukemia and gastrointestinal tumors. Several of our top predictions include those where Imatinib mesilate interacts with a number of other related tyrosine kinases,

including protein tyrosine kinase 6 (PTK6) and B-lymphoid tyrosine kinase both of which are either confirmed or candidate oncogenes.

4.4.2 Predicted GPCRs interaction network In the predicted GPCRs interaction network, there are some network components with respect to the GPCR families such as adrenergic receptor, purinergic receptor, cholinergic receptor, histamine receptor and dopamine receptor. β2-adrenergic receptor, for instance, interacts with more than 30 drugs in the gold standard dataset, and more than 100 ligands are predicted to interact with β2-adrenergic receptor. Opioid receptor is also known to interact with a wide variety of analgesics, and more than 30 derivatives are predicted to interact with opioid receptor. It is found that the drugs and compounds predicted by our method are chemically similar to the gold standard drugs and some of them are known analgesic agents.

Some GPCR families such as adrenergic receptor tend to have their members (α1, α2 and β2) clustered together because they share common ligands with each other. In the α2-adrenergic receptor network, predicted ligands like tiamenidine (D06125) are linked with all receptor nodes (α2a, α2b and α2c), while ligands like nisbuterol mesylate (D05171) are preferably predicted for α2a-adrenergic receptor. In the dopamine receptor network, many ligands are preferably predicted for dopamine receptor D2, and small number of ligands like perphenazine hydrochloride (D04965) is common among all dopamine receptors (D1, D2 and D3). The number of common ligands between dopamine receptors D1 and D2 is larger than that between dopamine receptors D1 and D3, which might reflect the similarities between dopamine receptor families.

5 DISCUSSION AND CONCLUSION

In this article, we characterized four classes of drug–target interaction networks in humans involving enzymes, ion channels, GPCRs and nuclear receptors, and revealed significant correlations between the drug structure similarity, the target sequence similarity and the drug–target interaction network topology. We then developed new statistical methods to predict unknown drug–target interaction networks from chemical structure information and genomic sequence information simultaneously on a large scale. The originality of the proposed method lies in the formalization of the drug–target interaction inference as a supervised learning problem for a bipartite graph, the lack of need for 3D structure information of the target proteins, and in the integration of chemical and genomic spaces into a unified space that we call 'pharmacological space'. In the results, we demonstrate the usefulness of our proposed method for the prediction of the four classes of drug–target interaction networks.

To date, there have been two research directions toward the detection of interactions between drug candidate compounds and target candidate proteins: the traditional drug discovery approach and the chemical biology approach. In the traditional drug discovery approach, we attempt to find new drug candidate compounds (or drug lead compounds) for a few certain proteins of interest. On the other hand, in the chemical biology approach, we attempt to find new target candidate proteins for a few certain chemical compounds of interest. Our proposed method has the advantages of both of the above approaches by finding new target candidate proteins and new drug candidate compounds simultaneously. It should be also pointed out that our proposed method can predict the interaction between previously unseen target candidate proteins and previously unseen drug candidate compounds which other methods including the nearest profile and weighted profile methods cannot.

A key observation is that two compounds sharing high structure similarity tend to interact with similar target proteins and hence are close in the network. Likewise two proteins sharing high sequence similarity tend to interact with similar drugs. However, there were some exceptional examples where this tendency was weak. For example, in the case of enzymes there exist many target proteins which share low sequence similarity but bind to similar drugs. This is reflected by the observation that the nearest profile and weighted profile methods often fail to predict the correct interaction pairs, because they are based on the direct use of sequence and chemical structure similarities. In contrast, our graph learning method is able to correct such biases, which is made possible by learning a model based on the partially known drug–target interaction network topology. It means that feature-based compound–protein pair score is inversely proportional to the network distance in the pharmacological feature space.

A variety of computational methods have been developed to analyze drug–target or compound–protein interactions. A powerful method is docking simulation (Cheng *et al.*, 2007; Rarey *et al.*, 1996), but it requires 3D structure information for the target proteins. Most pharmaceutically useful target proteins are membrane proteins such as ion channels and GPCRs. Determining the 3D structures of membrane proteins is still quite difficult which limits the use of docking. Our method does not need 3D structure information, but only the chemical structure information of the compounds and the sequence information of the proteins. Therefore, an advantage of our method is that it is suitable for screening a huge number of drug candidate compounds and target proteins on a large scale.

One previous research related with this study is the classification of target protein families based on the structure of their ligands (Keiser *et al.*, 2007). However, sequence information was not taken into consideration, and newly detectable interactions were limited to the linkage between known ligands and different protein families. The most recent work related with this study is the analysis of a global drug–target network consisting of different protein classes with a bipartite graph representation (Yildirim *et al.*, 2007), but the authors do not discuss the relationship with either protein sequence information or chemical structure information. On the other hand, we characterized four classes of drug–target interaction networks separately to examine the network features for each protein class, and revealed significant correlations between the target sequence similarity, drug structure similarity and the drug–target interaction network topology, which leads to the development of the methods to predict unknown drug–target interactions.

From a technical viewpoint, the performance of our method could be improved by using more sophisticated kernel similarity functions designed for genomic sequences and chemical structures (Schölkopf *et al.*, 2004). The incorporation of information about the functional sites into the protein similarity design is an interesting research direction (Kratochwil *et al.*, 2005). Recently, several kernel-based supervised network inference methods have been developed (Vert and Yamanishi, 2005; Yamanishi *et al.*, 2004), but they are limited to interactions between homogeneous molecules (e.g. protein–protein interactions) with a simple graph representation. In this study, we addressed the problem of predicting interactions between

heterogeneous molecules by regarding the interaction network as a bipartite graph. To our knowledge, there are no statistical methods to predict bipartite graphs in a supervised context. Our method can be applied to other biological network prediction problems such as metabolic network reconstruction and host–pathogen protein–protein interaction prediction as soon as they are represented by bipartite graphs.

In the final part of this article, we predicted interactions between all possible target candidate proteins and drug candidate compounds. Our comprehensively predicted drug–target interaction networks enable us to suggest many potential drug–target interactions. We confirmed that some of the interactions detected by our method corresponded to experimentally verified results in the literature. To detect new biological findings and potentially useful drug leads, we are currently working with collaborators on binding assays. We believe that our method is able to increase research productivity toward genomic drug discovery.

ACKNOWLEDGEMENTS

Funding: This work was supported by the 21st Century COE program 'Genome Science' from the Ministry of Education, Culture, Sports, Science and Technology of Japan, the Institute for Bioinformatics Research and Development of the Japan Science and Technology Agency, and the Japan Society for the Promotion Science. The Computational resource was provided by the Bioinformatics Center, Institute for Chemical Research and the Super Computer Laboratory, Kyoto University and the Super Computer System, Human Genome Center, Institute of Medical Science, The University of Tokyo.

Conflict of Interest: none declared.

REFERENCES

Cheng,A.C. *et al.* (2007) Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.*, **25**, 71–75.

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

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.

Haggarty,S.J. *et al.* (2003) Multidimensional chemical genetic analysis of diversity-oriented synthesis-derived deacetylase inhibitors using cell-based assays. *Chem. Biol.*, **10**, 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**, 11853–11865.

Kanehisa,M. *et al.* (2006) From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res.*, **34**(Database issue), D354–D357.

Keiser,M.J. *et al.* (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.*, **25**, 197–206.

Kratochwil,N.A. *et al.* (2005) An automated system for the analysis of g protein-coupled receptor transmembrane binding pockets: alignment, receptor-based pharmacophores, and their application. *J. Chem. Inf. Model.*, **45**, 1324–1336.

Kuruvilla,F.G. *et al.* (2002) Dissecting glucose signalling with diversity-oriented synthesis and small-molecule microarrays. *Nature*, **416**, 653–657.

Rainsford,K.D. (2007) Anti-inflammatory drugs in the 21st century. *Subcell. Biochem.*, **42**, 3–27.

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

Scholkopf,B. *et al.* (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, **10**, 1299–1319.

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–D433.

Smith,T.F. and Waterman,M. (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.

Vert,J.-P. and Yamanishi,Y. (2005) Supervised graph inference. *Adv. Neural Inf. Process. Syst.*, **17**, 1433–1440.

Wheeler,D.L. *et al.* (2006) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **34**, D173–D180.

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

Yamanishi,Y. *et al.* (2004) Protein network inference from multiple genomic data: a supervised approach. *Bioinformatics*, **20** (Suppl 1), i363–i370.

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

Zhu,S. *et al.* (2005) A probabilistic model for mining implicit 'chemical compound-gene' relations from literature. *Bioinformatics*, **21** (Suppl 2), ii245–ii251.