Relating drug–protein interaction network with drug side effects
Sayaka Mizutani¹, Edouard Pauwels²,³,⁴, Véronique Stoven²,³,⁴, Susumu Goto¹ and Yoshihiro Yamanishi¹,⁵,*
¹Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho Uji, Kyoto 611-0011, Japan,
²Mines ParisTech, Centre for Computational Biology, 35 rue Saint-Honoré, F-77305 Fontainebleau Cedex, France,
³Institut Curie, F-75248, Paris, France, ⁴INSERM U900, F-75248 Paris, France and ⁵Division of System Cohort,
Medical Institute of Bioregulation, Kyushu University, 3-1-1 Maidashi, Higashi-ku, Fukuoka, Fukuoka 812-8582, Japan

ABSTRACT
Motivation: Identifying the emergence and underlying mechanisms of drug side effects is a challenging task in the drug development process. This underscores the importance of system-wide approaches for linking different scales of drug actions; namely drug–protein interactions (molecular scale) and side effects (phenotypic scale) toward side effect prediction for uncharacterized drugs. Results: We performed a large-scale analysis to extract correlated sets of targeted proteins and side effects, based on the co-occurrence of drugs in protein-binding profiles and side effect profiles, using sparse canonical correlation analysis. The analysis of 658 drugs with the two profiles for 1368 proteins and 1339 side effects led to the extraction of 80 correlated sets. Enrichment analyses using KEGG and Gene Ontology showed that most of the correlated sets were significantly enriched with proteins that are involved in the same biological pathways, even if their molecular functions are different. This allowed for a biologically relevant interpretation regarding the relationship between drug-targeted proteins and side effects. The extracted side effects can be regarded as possible phenotypic outcomes by drugs targeting the proteins that appear in the same correlated set. The proposed method is expected to be useful for predicting potential side effects of new drug candidate compounds based on their protein-binding profiles.

Supplementary information: Datasets and all results are available at http://web.kuicr.kyoto-u.ac.jp/supp/mizutan/target-effect/
Availability: Software is available at the supplementary website.
Contact: yamanishi@bioreg.kyushu-u.ac.jp, or goto@kuicr.kyoto-u.ac.jp

1 INTRODUCTION
Predicting and countering the side effects of a new drug during its developmental phase remain important to the drug’s overall commercial success. Side effects are responsible for a significant number of cases where premarketed drugs fail during clinical trials. Identifying the underlying mechanisms of side effects is a challenging task, often because of the drugs’ pleiotropic effects on a biological system. Most drugs are small compounds that target and interact with proteins to induce perturbations in the proteins network. This underscores the need of system-wide approaches for predicting drug side effects by linking different scales of drug actions; drug–protein interactions (molecular scale) and relationships between drugs and side effects (phenotypic scale). Fliri et al. 2004, 2005, Tattonetti et al. 2004

The most widely used approach to identify possible side effects for a drug is to use its chemical structure information, based on the observation that drug chemical structures can reflect the ligand promiscuity toward protein targets (Rendler et al. 2010). For example, Scheiber et al. 2009 investigated correlations between drug chemical substructures and side effects, although they do not provide any prediction frameworks for new drug molecules. Yamanishi et al. 2012 proposed a method to predict the pharmacological effects of drugs using their chemical structures. They then inferred drug-target interactions, but their method cannot be directly applied to the prediction of high-dimensional side effect profiles. Atlas and Sharir 2013 proposed a method to predict side effects from chemical structure data using canonical correlation analysis (CCA). This work was pioneering in terms of simultaneous prediction of many side effects. Pauwels et al. 2011 proposed a method to relate drug chemical fragments with side effects using sparse CCA (SCCA), and used the chemical fragments to predict side effect profiles. However, these chemical structure-based methods cannot provide any biological interpretations regarding the underlying mechanisms at a molecular interaction level.

Chemically unrelated drugs may present similar side effects because they happen to share common off-target proteins Flinavonsa et al. 2004. On the basis of this observation, Campillos et al. 2008 used side effect similarity of marketed drugs to predict drug pairs with common protein targets. Xie et al. 2009 identified drug off-targets by docking the drug into protein binding pockets similar to that of its primary target, followed by mapping the proteins with the best docking scores to known biological pathways, thus predicting potential side effects. Using a similar docking approach, Wallach et al. 2010 searched for correlated pairs of side effects and biological pathways. These docking-based methods depend heavily on the availability of protein 3D structures, which presents serious limitations as many useful drug targets are membrane proteins, for which very few structures are available.

From a system-wide viewpoint, Fliri et al. 2008 performed a biological spectra-based approach to investigate the correlation between drug-targeted proteins and their side effects. They clustered drugs based on their biological spectra (i.e. their ability to inhibit each of 92 selected proteins) and revealed a correlation between the chemical structures of the corresponding drugs and their biological activity in terms of protein inhibition profile. They further showed that drugs with similar protein inhibition profiles tend to cause similar side effects Fliri et al. 2007. However, it remains difficult to predict the side effects of a drug based on its chemical structure.

© The Author(s) 2012. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/3.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

doi:10.1093/bioinformatics/bts383
to experimentally determine the link between drug-targeted proteins and side effects in a large-scale datasets in a cost-effective and efficient manner [Whitebread et al. 2005]. Therefore, there is a strong incentive to develop computational approaches for analyzing and predicting drug side effects.

In this article, we examine the correlation between drug–protein interactions and their side effects on a large scale, without limiting ourselves to proteins of known 3D structures. We identify correlated sets of proteins and side effects based on the co-occurrence of drugs in protein-binding profiles and in side effect profiles using SCCA. Results demonstrate that proteins in the same correlated set tend to be involved in only a few biological pathways even if their molecular functions are different. We also address that the side effects in each correlated set present possible outcomes from drug perturbations of corresponding proteins. The originality of the proposed method lies in the integration of drug–protein interactions at a molecular scale and drug side effect relationships at a phenotypic scale. Performance evaluation showed that this method works better than the case where drug side effect relationships are used in the SCCA framework. We also conduct a comprehensive side effect prediction for drug molecules stored in DrugBank without side effect information and confirm interesting predictions using independent source of information.

2 MATERIALS

2.1 Drug–protein interaction profiles and side effect profiles

Drug–protein interactions were obtained from DrugBank [Wishart et al. 2008] and Metador [Gläsinher et al. 2008]. Both the primary target proteins as well as all proteins known to directly interact with a particular drug were used for analysis. Side effect information was obtained from SIDER, which accumulates reported side effects from package inserts for marketed drugs [Kuhn et al. 2010]. In total, 658 drugs had both targeted protein and side effect information. This led to the construction of 5074 drug–protein interactions containing 1368 targeted proteins and 49,051 drug side effect pairs containing 1339 side effects. Each of the 658 drugs was represented by a 1368-dimensional protein-binding profile and 1339-dimensional side effect profile, which encodes for the presence or absence of proteins (side effects) by 1 or 0, respectively.

2.2 Chemical structures

To encode drug chemical structures, a fingerprint was used, which consisted of 881 chemical substructures defined in the PubChem database [R. et al 2010]. This resulted in a binary profile referred to as chemical substructure profile.

2.3 Annotation of drug-targeted proteins

265 pathway maps and 546 BRITE terms were obtained from Kyoto Encyclopedia of Genes and Genomes (KEGG) [Kanehisa et al. 2011]. Note that ‘global pathways’ and ‘disease pathways’ were excluded from the pathway set. Fourteen molecular function categories used in the protein annotations were also obtained from KEGG 22243 biological process terms and 9971 molecular function terms were obtained from Gene Ontology (GO) [Ashburner et al. 2000].

3 METHODS

3.1 Ordinary canonical correlation analysis (OCCA)

Suppose that we have a set of n drugs with p targeted protein features and q side effect features. Each drug is represented by a targeted protein feature vector $x=(x_1, \ldots, x_p)^T$, and by a side effect feature vector $y=(y_1, \ldots, y_q)^T$.

We consider two linear combinations for targeted proteins and side effects as $a_i = \alpha_i x_i$ and $b_i = \beta_i y_i$ ($i=1,2,\ldots,n$), where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)^T$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_n)^T$ are weight vectors. We attempt to find weight vectors $\alpha$ and $\beta$ which maximize the following canonical correlation coefficient:

$$\rho = \text{corr}(\alpha, \beta) = \frac{\sum_{i=1}^{n} \alpha_i \beta_i}{\sqrt{\sum_{i=1}^{n} \alpha_i^2} \sqrt{\sum_{i=1}^{n} \beta_i^2}}$$

where $\alpha$ (respectively, $\beta$) is centered. $\alpha$ (respectively, $\beta$) is called ‘canonical component’, score.

Let $X$ denote the $n \times p$ matrix defined as $X=[x_1, \ldots, x_n]^T$, and $Y$ denote the $n \times q$ matrix defined as $Y=[y_1, \ldots, y_q]^T$.

Then the maximization problem can be written as follows:

$$\max \|\alpha^T X^T \beta\|_2^2 \text{ subject to } \|\alpha\|_1 \leq 1, \|\beta\|_1 \leq 1.$$  

3.2 Sparse canonical correlation analysis (SCCA)

Most elements in the weight vectors $\alpha$ and $\beta$ in OCCA are non-zeros, which makes it difficult to interpret the result. In practice, it is desirable to find weight vectors that have large correlation, but that are also sparse for easier interpretation.

To impose the sparsity on $\alpha$ and $\beta$, we consider the following maximization problem with additional $L_1$ penalty terms:

$$\max \|\alpha^T X^T \beta\|_2^2 \text{ subject to } \|\alpha\|_1 \leq 1, \|\beta\|_1 \leq 1, \|\alpha\|_1 \leq c_1 \sqrt{p}, \|\beta\|_1 \leq c_2 \sqrt{q}.$$  

where $\|.|_1$ is $L_1$ norm (the sum of all absolute values of the vector elements), and $c_1$ and $c_2$ are parameters to control the sparsity ($0 < c_1 \leq 1$ and $0 < c_2 \leq 1$).

For simplicity, the same value is used for $c_1$ and $c_2$ in this study. TheCCA with $L_1$ penalties is referred to as SCCA. The weight vectors $\alpha$ and $\beta$ can be optimized by solving penalized matrix decomposition of the matrix $X = \Phi Y$ [Witten et al. 2009].

To obtain multiple canonical components, we perform a deflation manipulation iteratively as follows: $x^{k+1} = x^{k} - \lambda^k \alpha^T x^{k} \beta$ and $y^{k+1} = y^{k} - \lambda^k \beta^T y^{k} \alpha$, where $\lambda^k$ is the input of step $k$ ($\lambda^{k+1} = \lambda^k + \alpha^T x^{k} \beta$), $\alpha_k$ and $\beta_k$ are the weight vectors, and $d_k$ is singular value obtained in the $k$-th step (corresponding to the $k$-th component) ($k=1,2,\ldots,m$). Finally, we obtain $m$ pairs of weight vectors $(\alpha, \beta)$ and $m$ side effect profiles with non-zero weights in the weight vectors are extracted as correlated sets.

3.3 Prediction of side effect profiles for new molecules

Given the profile of targeted proteins $x_{new}$ for a drug of unknown side effects, we consider predicting its potential side effect profile $y_{new}$ based on the weight vectors $[\alpha_1, \ldots, \alpha_m]$ and $[\beta_1, \ldots, \beta_m]$.

We use the following prediction score for a given molecule:

$$y_{new} = \sum_{k=1}^{m} \beta_k \alpha_k^T x_{new} = R L A^T x_{new},$$

where $A = [\alpha_1, \ldots, \alpha_m]$, $R = [\beta_1, \ldots, \beta_m]$, and $L$ is the diagonal matrix whose elements are canonical correlation coefficients. If the $j$-th element in $y_{new}$ has a high score, the new molecule is predicted to have the $j$-th side effect ($j=1,2,\ldots,q$). The same prediction score was proposed in the previous work [Parsons et al. 2013].

3.4 Enrichment analyses of targeted proteins

Let $G_i$ denote the set of extracted proteins in component $i$ and $G$ denote the set of proteins in a functional unit (e.g. KEGG pathway map). Let
We applied the proposed SCCA method to the protein-binding dataset. We assume that the probability to observe an intersection of size $z$ between $G$ and $G_G$ follows a hypergeometric distribution. The enrichment score $s(c)$ of a component $c$ is computed as follows:

$$p(G_G) = \frac{\sum_{i=0}^{r} \binom{z}{i} \binom{r-i}{k-i}}{\binom{r}{k}},$$ (5)

We then define the enrichment score $s(c)$ of a component $c$ by

$$s(c) = -\log_{10}(p(G_G))$$

where $p^{FDR}(G_G)$ is the corrected value of $p(G_G)$ by the false discovery rate (FDR) (Benjamini and Hochberg, 1995).

4 RESULTS

4.1 Extraction of canonical component sets of drug-targeted proteins and side effects

We applied the proposed SCCA method to the protein-binding profiles and side effect profiles (see ‘Materials’ section), which provided us with 80 canonical components. The correlated sets of proteins and side effects were extracted from each component. We refer to these correlated sets as canonical components (CCs) hereafter. All components present a limited number of proteins and side effects, which is a consequence of the sparsity of SCCA. This allows meaningful biological interpretation, indicating an advantage over OCCA.

Figure 1 illustrates the network of extracted targeted proteins and side effects by connected diamonds if they appear in the same canonical component (CC). The highlighted components (CCs) hereafter. All components present a limited number of proteins and side effects, which is a consequence of the sparsity of SCCA. This allows meaningful biological interpretation, indicating an advantage over OCCA.

Figure 1 shows the distributions of canonical components against the number of enriched functional units associated with the component(s). The results of pathway enrichment analyses (Fig. 1 and b) displayed skewed distributions, i.e. the number of components decreased as the number of enriched pathways increased. Pathway enrichment analysis with KEGG pathway maps showed that 33 components were enriched with one or two pathway(s), and the additional 23 components were enriched with <10 pathways. This trend suggests that the majority of components were characterized by only a small number of KEGG pathways. Pathway enrichment analysis with GO biological process terms also displayed this trend. In contrast, the component distributions for molecular functions in Figure 1d showed much less skewness.
Table 1. Statistics in pathway and molecular function enrichment analyses

<table>
<thead>
<tr>
<th></th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number annotated proteins</td>
<td>215</td>
<td>281</td>
<td>298</td>
<td>298</td>
</tr>
<tr>
<td>Number of pathways/terms used in annotation</td>
<td>112</td>
<td>751</td>
<td>105</td>
<td>318</td>
</tr>
<tr>
<td>Number of components with enrichment</td>
<td>57</td>
<td>72</td>
<td>75</td>
<td>74</td>
</tr>
<tr>
<td>Number of enriched pathways/terms</td>
<td>33</td>
<td>93</td>
<td>50</td>
<td>75</td>
</tr>
</tbody>
</table>

Four types of functional units were tested: (a) KEGG pathway maps; (b) GO biological process; (c) KEGG BRITE terms and (d) GO molecular function.

and components were distributed across varying numbers of terms. This suggests that components are likely to represent a few distinct biological pathways for proteins from different molecular functions.

However, there remains a possibility that molecular functions of similar categories appear in a component. Thus, we also examined molecular functions of proteins in more general categories, such as ‘Ion channels’, ‘G Protein-coupled receptors’, and ‘Enzymes’. For the majority of components, as a result, proteins were annotated with more than one molecular function category, which confirms that proteins of very different molecular functions were grouped into the same component.

Supplementary Fig. S1)

Table 2 shows the 10 most frequently appearing enriched KEGG pathway maps that showed enrichment. Component-based enriched pathways and associated enrichment scores are shown in Supplementary Table S2. Pathways co-appearing in the same component are biologically relevant. Among the 40 components that are enriched with more than one pathway, 14 contained two or three of the following pathways: ‘Calcium signaling pathway’, ‘Cardiac muscle contraction’, and ‘MAPK signaling pathway’. Interestingly, calcium is a signaling molecule that is well known to play an important role in muscle contraction, and there exists a direct cause-and-effect relationship between MAP kinase activation and smooth muscle contraction (Kesy et al., 1999). Therefore, in many cases where components contain more than one pathway, these pathways seem to take part in the same global biological function.

We are aware that the KEGG pathways do not always describe signal transductions of proteins. For example, ‘Neuroactive ligand–receptor interaction pathway’ lists ligand–receptor interactions for G protein-coupled receptors (GPCRs) and ion channels. It describes protein families, rather than signal transductions occurring downstream of the ligand–receptor interactions. However, 18 out of 19 components were enriched with additional pathways.

These results indicate that component-based grouping of targeted proteins provides biologically relevant information in two ways. First, a significant number of proteins co-extracted in the same component are involved in the same biological pathway(s). Second, in many components, such proteins belong to different protein families. The component-based protein grouping cannot be inferred only from drug-targeted protein interactions, because they are often targeted by drugs of different chemical families. Accordingly, the side effects extracted in a component can be viewed as possible outcomes of biological pathway perturbations by drugs targeting the proteins that appear in the component.

4.3 Performance evaluation

It is difficult to evaluate the performance of the feature extraction method, because there is little knowledge about true association between targeted proteins and side effects. However, if the extracted components are biologically meaningful, they should contain some general properties which could be exploited for side effect prediction. We evaluate the performance of the method by recovering known drug side effect profiles from drug protein-binding profiles, using the extracted canonical components.

In previous literature, chemical structure fingerprints were used for predicting side effect profiles in the framework of OCCA (Atias and Sharan, 2011) and SCCA (Pauwels et al., 2007). Therefore, we made a comparison between chemical structure-based approach and targeted protein-based approach in the framework of both OCCA and SCCA by performing the following 5-fold cross-validation.

First, drugs in the gold standard set were split into five subsets of roughly equal sizes, and each subset was used in turn as a test set. Second, the CCA model was trained on the remaining four sets. Third, the prediction score was computed from the test set, based on the components extracted from the training set. Finally, the model was evaluated for prediction accuracy over the 5-folds.

We evaluated the performance of the methods by the receiver-operating characteristic curve (ROC curve) and the Precision–Recall curve (Supplementary Fig. S2). The ROC curve is a plot of true positives as a function of false positives based on various prediction score thresholds, where true positives are correctly predicted side effects and false positives are incorrectly predicted side effects. The Precision–Recall curve is the plot of ‘precision’ (positive
predictive value) as a function of ‘recall’ (sensitivity) based on various thresholds.

We summarized the performance by the area under the ROC curve (AUC) score and the area under the Precision-Recall curve (AUPR) score. To obtain robust results, we repeated the cross-validation experiment five times, and computed the mean and the standard deviation (SD) of the AUC scores over the five repetitions. Sparsity parameters $c_1$, $c_2$ ranged from 0 to 1 by 0.1 increments, and $m$ ranged from 10 to 200 by 10 increments. The best results were obtained with $c_1 = 0.1$, $c_2 = 0.1$ and with $m = 80$ components in the case of SCCA. The same cross-validation experiments were repeated for OCCA (with no sparsity constraint), and the best results were obtained for $m = 20$ components.

Table 2 shows the resulting AUC and AUPR scores for the four different approaches, where the prediction scores for all side effects were merged, and a global ROC curve and a global PR curve were evaluated for each approach. This indicates that both SCCA and OCCA produce fairly good results and SCCA is slightly better than OCCA. It also seems that the targeted protein-based approach works better than the chemical structure-based approach. Results suggest that the targeted protein information is indeed useful for side effect prediction.

### 4.4 Prediction of side effects for uncharacterized drugs

In the DrugBank database, there are still 730 drugs whose target protein information is available, but side effects are not stored in the SIDER database. On the basis of their protein-binding profiles, we predicted the potential side effects for these uncharacterized drugs using the SCCA model, all of the 658 reference drugs being used as a training set. All prediction results can be found in Supplementary Table S3A. Complete analysis of all predictions is out of course out of reach, so we focused on the side effect predictions of highest scores.

Some of the top-ranked predicted side effects involve Cinnarizine (DB00558), an anti-histaminic drug used against motion sickness. This drug binds to the histaminic H1 receptor, which is believed to explain its effectiveness in preventing vomiting in motion sickness. Its predicted ‘tremor’ (cyclical movement of a body part) side effect was confirmed by literature [Kaminew-Roklan and Mäurer 1993]. Interestingly, Cinnarizine also binds to the voltage-dependent calcium channel involved in muscle contraction, which might explain the ‘tremor’ side effect. ‘Constipation’ is also predicted for Cinnarizine, as found in the adverse effect report 6127929-0 of the Food and Drug Administration (FDA). The predicted ‘dry mouth’ side effect for Cinnarizine was also confirmed from literature [Gordon et al. 2003].

The second ranked predicted side effect is ‘diplopia’ (double vision) for Benzocaine (DB01086), a surface anesthetic that acts by preventing transmission of impulse along nerve fibers. Consistent with this activity, Benzocaine is an inhibitor of voltage-dependent sodium channel. The predicted side effect was confirmed from the literature [Horowitz et al. 2001]. ‘Syncope’ was another side effect predicted for Benzocaine with a high score [Walker et al. 2001]. The fourth ranked predicted side effect is ‘tremor’ for Bepridil (DB01244), an antihypertensive drug. Tremor is indeed one of the most common side effects for this drug, reported for 5% of all patients [Williams et al. 2006]. ‘Tachycardia’ was also confirmed for Promazine (DB00420), an antipsychotic agent [Kronenberg 2004].

The side effect ‘diplopia’ for Nisoldipine (DB00401), a calcium channel blocker used for the management of hypertension, was also confirmed [Keefe and Cramer 1993].

### 5 DISCUSSION

We provide biological interpretations of proteins and side effects extracted in each canonical component CC. Although CCs are discussed here, the distinct characteristics were also observed in many other components.

In CC1, the top-ranked proteins were serotonin and noradrenaline transporters. They belong to the family of neurotransmitter transporters and are responsible for the release and re-uptake of the serotonin and noradrenaline neurotransmitter molecules by neurons, at the level of synapses. The uptake of the neurotransmitters is coupled to the co-transport of sodium ion by sodium channels to derive the required energy to pump the neurotransmitter against its gradient. Therefore, neurotransmitter transporters and sodium channels can be viewed as proteins playing part in the same biological process. This indicates that drugs targeting sodium channels or neurotransmitter transporters share some side effects, indeed, drugs such as Venlafaxine or Bupropion that target neurotransmitter transporters gained high score in CC1, together with drugs such as Plecanide or Lamotrigine that target sodium channels. As shown in Figure 2a and b, channel blockers and drugs targeting serotonin and noradrenaline transporters display very diverse chemical structures, because they probably target sodium channels.

**Table 2.** Most frequently appearing enriched pathways

<table>
<thead>
<tr>
<th>ID</th>
<th>KEGG pathway maps</th>
</tr>
</thead>
<tbody>
<tr>
<td>map04080</td>
<td>Neuronal network signaling</td>
</tr>
<tr>
<td>map04020</td>
<td>Calcium signaling pathway</td>
</tr>
<tr>
<td>map04728</td>
<td>Dopaminergic synapse</td>
</tr>
<tr>
<td>map04010</td>
<td>MAPK signaling pathway</td>
</tr>
<tr>
<td>map04260</td>
<td>Cardiac muscle contraction</td>
</tr>
<tr>
<td>map04727</td>
<td>GABAergic synapse</td>
</tr>
<tr>
<td>map04970</td>
<td>Salivary secretion</td>
</tr>
<tr>
<td>map04725</td>
<td>Cholinergic synapse</td>
</tr>
<tr>
<td>map04090</td>
<td>Arachidonic acid metabolism</td>
</tr>
<tr>
<td>map04270</td>
<td>Vascular smooth muscle contraction</td>
</tr>
</tbody>
</table>

**Table 3.** Performance evaluation based on 5-fold cross-validation

<table>
<thead>
<tr>
<th>Method</th>
<th>AUC ± S.D.</th>
<th>AUPR ± S.D.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chemical structure-based approach</td>
<td>Random: 0.5000 ± 0.0000, OC: 0.8355 ± 0.0010</td>
<td>SCCA: 0.8708 ± 0.0007, SCCA: 0.8708 ± 0.0007</td>
</tr>
<tr>
<td>OCCA</td>
<td>SCCA: 0.8708 ± 0.0007, SCCA: 0.8708 ± 0.0007</td>
<td></td>
</tr>
</tbody>
</table>

Scores of the proposed method are highlighted in bold.
very different binding sites. The fact that they lead to similar side effects could not have been foreseen based on analysis of chemical structures. Dopamine receptor was also found in CC1. This receptor belongs to the GPCR family and is implicated in anti-inflammatory proteins such as annexin 1, or down-regulates ligand complex translocates to the nucleus where it up-regulates receptors are activated by glucocorticoid binding. The receptor–inflammation down. In this pathway, cytosolic glucocorticoid of the feedback mechanism in the immune system that turns glucocorticoid signaling pathway, such as glucocorticoid effects and contributed to the same component. Venlafaxine has a high score in CC1 and is an example of such a case.

In CC2, the top-ranked proteins were gamma-aminobutyric acid (GABA) receptors (GABA-A receptors). These receptors function as chloride channels gated by GABA. They mediate neuronal inhibition in the central nervous system. Consistent with this role, most of the high scoring drugs in CC2 were molecules that target GABA receptors and modulate their function. They are used as anti-anxiety agents, muscle relaxants or anesthesia adjuvants. Midazolam and Baclofen are examples of drugs targeting GABA receptors (Figure B), although these molecules belong to very different chemical families. Midazolam is a benzodiazepine molecule that binds to GABA-A receptors and acts as an agonist that increases GABA activity. Baclofen is a GABA derivative that targets GABA type B receptors (which are GPCRs) in addition to the GABA-A receptors. Interestingly, Gabapentin was one of the high scoring drugs in CC2 (Figure B), although it does not bind to GABA receptors. This molecule binds to and inhibits voltage-sensitive calcium channels. However, this drug is known to increase GABA concentration in the synapse, although the corresponding mechanism is not understood (Petroff et al 2000). Consequently, it indirectly increases the activity of GABA receptors, which is consistent with its use (among others) as an anti-anxiety agent.

Proteins extracted in CC5 and CC15 are mainly involved in pathways related to inflammation. In CC5, most proteins with high weights are enzymes belonging to the arachidonic acid metabolism pathway, such as prostaglandin G/H synthases, arachidonate 5-lipoxygenase and leukotriene A-4-hydroxylase. Indeed, CC5 was enriched with the KEGG ‘Arachidonic acid metabolism’ pathway, in which proinflammation molecules such as leukotrienes or prostaglandins are synthesized. High scoring drugs in CC5 were mainly non-steroidal anti-inflammatory drugs (NSAIDs), that inhibit enzymes of this pathway. Some of these drugs are structurally unrelated, such as Diclofenac and Indomethacin, but they both inhibit prostaglandin G/H synthases (Figure B). Others also have very different chemical structures, such as Leflunomide that binds arachidonate 5-lipoxygenase (Figure B). However, all these drugs exert their anti-inflammatory action by inhibiting the same overall biological pathway, which explains why they lead to common side effects and contributed to the same component.

In CC15, most of the extracted proteins belong to the glucocorticoid signaling pathway, such as glucocorticoid receptor (a nuclear receptor) or corticosteroid-binding globulin. Glucocorticoids are a class of steroid hormones that are part of the feedback mechanism in the immune system that turns inflammation down. In this pathway, cytosolic glucocorticoid receptors are activated by glucocorticoid binding. The receptor–ligand complex translocates to the nucleus where it up-regulates anti-inflammatory proteins such as annexin 1, or down-regulates proinflammatory proteins such as interleukins or cytokines. High scoring drugs in CC15 were molecules from the steroid family such as Triamcinolone that binds glucocorticoid receptor or Aminocortide that binds glucocorticoid receptor and annexin 1 (Figure H). Therefore, although high scoring drugs in CC5 or CC15 present an anti-inflammation activity, they do not act on the same biological pathways. Consequently, they do not present the same side effects.

Interestingly, CC5 contained a human DNA topoisomerase. This protein controls the topological states of DNA, and therefore, one could wonder why this protein was extracted in CC5. In fact, various drugs from the fluoroquinolone family, namely, Ciprofloxacin, Ofloxacin and Doxorubicin, which target DNA topoisomerase were found among high scoring drugs in CC5. These fluoroquinolones and NSAIDs, or example Diclofenac and Indomethacin, share some of the extracted side effects of CC5, namely, ‘Stevens Johnson syndrome’, ‘ stomatitis’ and ‘agranulocytosis’. NSAIDs and fluoroquinolones are structurally unrelated molecules, have completely different modes of action and target functionally unrelated proteins. The fact that they share the common side effects may not be explained by our current pathway data. Additional information such as relationships between pathways will be required to fully explain such cases.

Another case is a possibility that certain side effects can be caused by an alteration of the immune system introduced by drugs. It is known that synthetic glucocorticoids down-regulate the functions of immune cells (Plummer and Regensiek 2001), and certain anti-cancer drugs show immunosuppressive activities (Lex 2005). To discuss whether the occurrence of common side effects between these drugs is due to the direct inhibitions of the targeted proteins or indirect alterations of the immune system, an investigation of cross talks between signaling pathways and immune pathways is necessary.
6 CONCLUSION

In this article, we proposed a novel SCCA-based approach to relate drug targeted proteins with drug side effects. Using a cross-validation scheme, we found that the proposed approach displays better performance than chemical-structure-based methods for the prediction of drug side effects. Results suggest that side effect of drugs are more correlated to their mechanism of action, rather than to their chemical structure, which presents an interesting result. In most drug discovery projects, a therapeutic target playing a role in a given disease is searched for, and once identified, the corresponding pathways can be identified. The components that are enriched in these pathways provide a list of potential side effects that one can expect for future drugs acting on the target of interest.

We constructed a statistical model for the prediction of side effect profiles from protein-binding profiles, primarily because the number of drugs with side effect information is much less than those with targeted protein information. Indeed, it is unlikely in a practical situation that detailed side effect profiles are known for newly arriving drug candidate molecules. One limitation of our proposed method is that targeted protein information is not always obtainable; however, increasing information regarding protein-ligand interactions is becoming available from various biological assays. Thus, we envisage scenarios where a drug candidate molecule’s targeted protein information is available, but not side effect information. In this context, we believe that our proposed method presents itself as a powerful and informative tool for use within the drug discovery process.

Funding: (in part) JSPS International Training Program (ITP) and the Japan Science and Technology Agency. Computational resources were provided by Bioinformatics Center, Institute for Chemical Research, Kyoto University.

Conflict of Interest: none declared.

REFERENCES