JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

Extracting Sets of Chemical Substructures and Protein Domains Governing Drug-Target Interactions

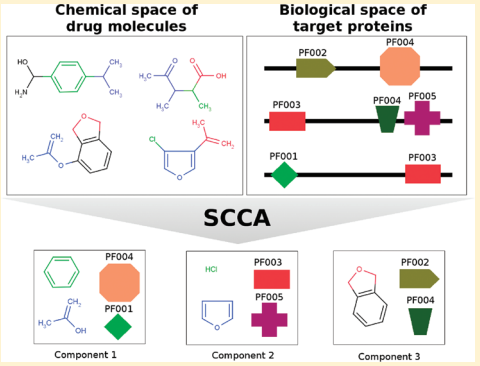
Yoshihiro Yamanishi,*[,†] Edouard Pauwels,[†] Hiroto Saigo,[‡] and Véronique Stoven[†]

[†]Mines ParisTech, Centre for Computational Biology, 35 rue Saint-Honore, F-77305 Fontainebleau Cedex, France, Institut Curie, F-75248, Paris, France, and INSERM U900, F-75248 Paris, France

[‡]Max-Planck-Institut für Informatik, Computational Biology and Applied Algorithmics, Campus E1 4, 66123 Saarbrücken, Germany

Supporting Information

ABSTRACT: The identification of rules governing molecular recognition between drug chemical substructures and protein functional sites is a challenging issue at many stages of the drug development process. In this paper we develop a novel method to extract sets of drug chemical substructures and protein domains that govern drug-target interactions on a genome-wide scale. This is made possible using sparse canonical correspondence analysis (SCCA) for analyzing drug substructure profiles and protein domain profiles simultaneously. The method does not depend on the availability of protein 3D structures. From a data set of known drug-target interactions including enzymes, ion channels, G protein-coupled receptors, and nuclear receptors, we extract a set of chemical substructures shared by drugs able to bind to a set of protein domains. These two sets of extracted chemical substructures and protein domains form components that can be further exploited in a drug discovery process. This approach successfully clusters protein domains that may be evolutionary unrelated but that bind a common set of chemical substructures. As shown in several examples, it can also be very helpful for predicting new protein−ligand interactions and addressing the problem of ligand specificity. The proposed method constitutes a contribution to the recent field of chemogenomics that aims to connect the chemical space with the biological space.



INTRODUCTION

Most drugs are small chemical compounds which interfere with the biological behavior of their target proteins; therefore, identification of interactions between ligand compounds and target proteins is a key area in drug discovery. A commonly used computational approach to analyze and predict ligand-protein interactions is docking (see ref 1 for a recent review), but docking cannot be applied to proteins with unknown 3D structures, which limits its use. This limitation is critical in the case of membrane proteins such as G protein-coupled receptors (GPCRs) or ion channels, which are major therapeutic targets, but whose 3D structure determination is known to be particularly difficult.

The importance of chemogenomic approach is growing fast in recent years,[2−4] and a variety of statistical methods based on chemical and genomic information have been proposed to predict drug-target or more generally ligand-protein interactions. These methods assume that similar proteins are expected to bind similar ligands. They differ by the underlying description used for proteins and ligands and by how similarities between these objects are measured. Examples are statistic-based methods that compare target proteins by the similarity of the ligands that bind to them, which can then be used to predict new protein−ligand interactions.[5,6] Other approaches are the binary classification approaches such as support vector machine with pairwise kernels for compound-protein pairs[7−9] and the supervised bipartite graph inference with distance learning based on chemical and genomic similarities.[10,11]

Ligand-protein interactions are often due to common chemical structures (the pharmacophore) that are usually shared by the ligands of a given protein, whereas this is not expected for random compounds that do not bind to the same protein. Recently, a variety of analyses have been conducted, such as analysis of chemical substructures and biological activity,[12] data mining of chemical structural fingerprints and high-throughput screening data in PubChem,[13] or extraction of chemical modification patterns in drug development.[14] Ligand-protein interactions are also due to functional sites of proteins (e.g., binding pockets, domains, motifs). Recently, the comparison of binding pockets has been done to investigate the relationship with their ligands.[15−17] However, these methods require the availability of the 3D structure of proteins. To date, most of the research has been performed separately from the viewpoints of either ligands or proteins. Yet, the relevant question is how to relate ligand chemical substructures with protein functional sites in terms of ligand-protein interactions. There is therefore a strong incentive to conduct an

Received: December 6, 2010
Published: April 21, 2011

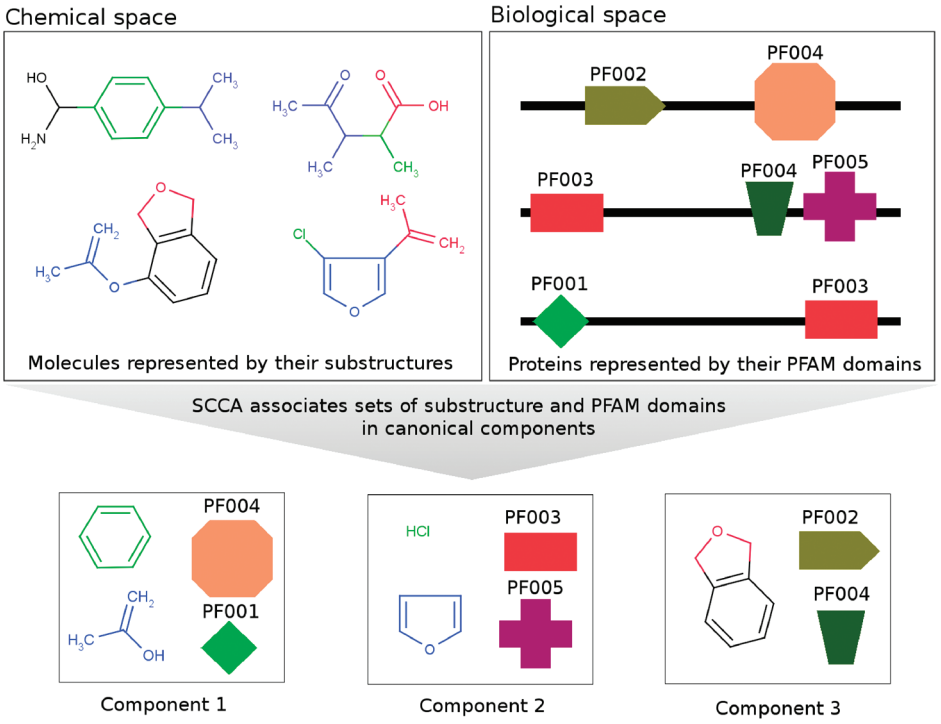


Figure 1. An illustration of the proposed method.

integrative analysis of both ligand substructures and protein functional sites toward understanding of ligand-protein interactions. In this domain, a challenging issue is to develop methods that identify rules for molecular recognition between ligand chemical substructures and protein functional sites.

In this paper we develop a novel method to extract sets of drug chemical substructures and protein domains that govern drug-target interactions. This is made possible using canonical correspondence analysis (CCA) for analyzing drug substructure profiles and protein domain profiles simultaneously. We develop an extension of the CCA algorithm by incorporating sparsity for easier interpretation, which we call sparse canonical correspondence analysis (SCCA). Figure 1 shows an illustration of the proposed method. The main interest and originality of the proposed method is that it correlates protein domains to chemical substructures expected to be present in their ligands, based on a learning data set. In other words, the method identifies pharmacophores automatically, explaining why a given molecule binds to a given protein domain. As we show in the following, the method may also help to address the question of specificity.

From a learning data set of drug-target interactions constructed from the DrugBank database, the proposed method extracts sets (called components) formed by a set of chemical substructures shared by drugs able to bind to a set of protein domains. The method successfully clusters protein domains that may be evolutionary unrelated, but whose ligands share a common set of chemical substructures. We show how the method can be a useful tool, among others, in the drug development process. This method constitutes a contribution to the recent field of chemogenomics that aims to connect the chemical space with the biological space.

MATERIALS

Drug-target interactions were obtained from the DrugBank database which combines detailed data about drugs and drug candidates with comprehensive drug-target information.[18] The version of DrugBank is 2.5. Proteins belong to many different classes, among others, pharmaceutically useful ones such as enzymes, ion channels, G protein-coupled receptors (GPCRs), or nuclear receptors. In this study, we focused on human proteins, which drove us to select all interactions involving human proteins. This led to build a protein-drug data set containing 4809 interactions involving 1554 proteins and 1862 drugs. The set of interactions is used as a gold standard data.

To encode the chemical structures of drugs involved in these interactions, we used a fingerprint corresponding to the 881 chemical substructures defined in the PubChem database.[19] Each drug was represented by an 881 dimensional binary vector whose elements encode for the presence or absence of each PubChem substructure by 1 or 0, respectively. Most of the drugs documented in DrugBank have a link to PubChem, but some do not, mainly biotech drugs and mixtures. Interactions involving drugs that have a record in PubChem were kept. Among the 881 substructures used to represent the chemical structures, 663 are actually used, because 218 do not appear in our drug set.

For all proteins, genomic information and annotation were obtained from the UniProt database,[20] and associated protein domains were obtained from the PFAM database.[21] In the set of proteins we took into account, 876 PFAM domains are found. Therefore, each protein was represented by an 876 dimensional binary vector whose elements encode for the presence or absence of each of the retained PFAM domain by 1 or 0, respectively.

METHODS

We want to extract drug chemical substructures and protein domains which tend to jointly appear in the interaction pairs of drugs and target proteins and to disappear in the other pairs. A possible statistical approach for achieving this goal is the canonical correspondence analysis (CCA). We first make a brief review

of ordinary CCA (OCCA), and we then develop a new method, which we call sparse CCA (SCCA).

Ordinary Canonical Correspondence Analysis (OCCA). Suppose that we have a set of n_x drugs with p substructure features, a set of n_y target proteins with q domain features, and information about interactions between the drug set and the target protein set. Note that $n_x \neq n_y$. Each drug is represented by a p-dimensional feature vector $\mathbf{x} = (x_1,\cdots,x_p)^T$, and each target protein is represented by a q-dimensional feature vector $\mathbf{y} = (y_1,\cdots,y_q)^T$.

Consider two linear combinations for drugs and proteins as $u_i = \alpha^T\mathbf{x}_i$ $(i = 1,2,\cdots,n_x)$ and $v_j = \beta^T\mathbf{y}_j$ $(j = 1,2,\cdots,n_y)$, repectively, where $\alpha = (\alpha_1,\cdots,\alpha_p)^T$ and $\beta = (\beta_1,\cdots,\beta_q)^T$ are weight vectors. The goal of ordinary CCA is to find weight vectors α and β which maximize the following canonical correlation coefficient

$$corr(u,v) = \frac{\sum_{i,j} I(\mathbf{x}_i,\mathbf{y}_j)\alpha^T\mathbf{x}_i\cdot\beta^T\mathbf{y}_j}{\sqrt{\sum_i d_{x_i}(\alpha^T\mathbf{x}_i)^2}\sqrt{\sum_j d_{y_j}(\beta^T\mathbf{y}_j)^2}} \quad (1)$$

where $I(\cdot,\cdot)$ is an indicator function which returns 1 if drug $\mathbf{x}_i$ and protein $\mathbf{y}_j$ interact with each other or 0 otherwise, d_{x_i} (respectively d_{y_j}) is the degree of $\mathbf{x}_i$ (respectively $\mathbf{y}_j$), $\Sigma_i u_i = 0$ (respectively $\Sigma_j v_j = 0$) is assumed, and u (respectively v) is called *canonical components* for $\mathbf{x}$ (respectively $\mathbf{y}$).[22] This maximization problem can be written as follows

$$\max\{\sum_{i,j} I(\mathbf{x}_i,\mathbf{y}_j)\alpha^T\mathbf{x}_i\cdot\beta^T\mathbf{y}_j\} \text{ subject to}$$
$$\sum_i d_{x_i}(\alpha^T\mathbf{x}_i)^2 \le 1,\ \sum_j d_{y_j}(\beta^T\mathbf{y}_j)^2 \le 1 \quad (2)$$

Here we define the $n_x \times n_y$ adjacency matrix A, where element $(A)_{ij}$ is equal to 1 (respectively 0) if drug $\mathbf{x}_i$ and protein $\mathbf{y}_j$ are connected (respectively disconnected). Let X be the $n_x \times p$ matrix defined as $X = [\mathbf{x}_1,\cdots,\mathbf{x}_{n_x}]^T$, and let Y denote the $n_y \times q$ matrix defined as $Y = [\mathbf{y}_1,\cdots,\mathbf{y}_{n_y}]^T$, where the columns of X and Y are assumed to be centered and scaled. In the matrix form, we can rewrite the above optimization problem as follows

$$\max\{\alpha^T X^T A Y\beta\} \text{ subject to}$$
$$\alpha^T X^T D_x X\alpha \le 1,\ \beta^T Y^T D_y Y\beta \le 1 \quad (3)$$

where D_x and D_y are matrices whose diagonals are the degrees of drugs and target proteins, respectively. In other high-dimensional problems, it is known that good results can be obtained by treating the covariance matrix as a diagonal matrix.[23,24] Therefore, we substitute identity matrices for $X^T D_x X$ and $Y^T D_y Y$ and consider the following optimization problem

$$\max\{\alpha^T X^T A Y\beta\} \text{ subject to } ||\alpha||_2^2 \le 1,\ ||\beta||_2^2 \le 1 \quad (4)$$

where $\|\cdot\|_2$ is L_2 norm (the square root of the sum of squared values in the vector).

Sparse Canonical Correspondence Analysis (SCCA). In the OCCA, the weight vectors α and β are not unique if p exceeds n_x or q exceeds n_y. In addition, it is difficult to interpret the results when there are many nonzero elements in the weight vectors α and β. In practical applications, especially when p and q are large, we want to find weight vectors α and β for $\mathbf{x}$ and $\mathbf{y}$ that have large correlation but are also sparse for easier interpretation.

To impose the sparsity on α and β, we propose to consider the following optimization problem with some additional L_1 penalty terms

$$\max\{\alpha^T X^T A Y\beta\} \text{ subject to}$$
$$||\alpha||_2^2 \le 1,\ ||\beta||_2^2 \le 1,\ ||\alpha||_1 \le c_1\sqrt{p},\ ||\beta||_1 \le c_2\sqrt{q} \quad (5)$$

where $\|\cdot\|_1$ is L_1 norm (the sum of absolute values in the vector), and c_1 and c_2 are parameters to control the sparsity and restricted to ranges $0 < c_1 \le 1$ and $0 < c_2 \le 1$. The sparse version of CCA is referred to as sparse canonical correspondence analysis (SCCA).

The optimization problem in SCCA can be regarded as the problem of penalized matrix decomposition of the matrix $Z = X^T A Y$. Recently, a useful algorithm for solving the penalized matrix decomposition (PMD) problem has been proposed.[25] In order to obtain the solutions of SCCA, we propose to apply the PMD algorithm to the matrix $Z = X^T A Y$.

Let S denote the soft-thresholding operator defined as $S(a,c) = sgn(a)(|a| - c)_+$ where c is a positive constant, and x_+ is defined to equal x if $x > 0$ and 0 otherwise.

Here the criterion (to be maximized) is denoted as $\rho = \alpha^T Z\beta$ and is referred to as the singular value.

Then, the PMD algorithm in this context is as follows:

1 set α and β to have L_2 norm 1.
2 $\alpha \leftarrow (S(Z\beta,\delta_1))/(\|S(Z\beta,\delta_1)\|_2)$, where $\delta_1 = 0$ if this results in $\|\alpha\|_1 \le c_1$: otherwise, δ_1 is chosen to be a positive constant such that $\|\alpha\|_1 = c_1$.
3 $\beta \leftarrow (S(Z^T\alpha,\delta_2))/(\|S(Z^T\alpha,\delta_2)\|_2)$, where $\delta_2 = 0$ if this results in $\|\beta\|_1 \le c_2$: otherwise, δ_2 is chosen to be a positive constant such that $\|\beta\|_1 = c_2$.
4 $\rho \leftarrow \alpha^T Z\beta$.
5 repeat steps 2, 3, and 4 until convergence.

Note that the above algorithm can be used for finding one canonical component. In order to obtain multiple canonical components, we propose to iterate the maximization of the above criterion repeatedly, each time using the Z matrix as the residuals obtained by subtracting from the matrix the previous factors found (deflation), that is, we recursively estimate the k-th weight vectors α_k and β_k for $k = 1,2,\cdots,m$ as follows:

1 set $Z^{(1)} \rightarrow Z$.
2 find α_k, β_k, and ρ_k by applying the above PMD algorithm to $Z^{(k)}$.
3 $Z^{(k+1)} \rightarrow Z^{(k)} - \rho_k\alpha_k\beta_k{}^T$.
4 repeat steps 2 and 3 for $k = 1,2,\cdots,m$.

Finally, we obtain m pairs of weight vectors $\alpha_1,\cdots,\alpha_m$ and $\beta_1,\cdots,\beta_m$. For easier interpretation, the sign of the weight vectors is adjusted such that the weight element with the highest absolute value is positive in each component.

Components of lower k are called "lower order components", while components of higher k are called "higher order components".

High scoring substructures and domains in the weight vectors are considered important in terms of drug-target interactions. We evaluate the strength of the association between chemical substructures and protein domains by computing the product between the weight elements in α and the weight elements in β within each component.

The originality of the SCCA method lies in the development of a sparse version of canonical correspondence analysis to handle the heterogeneous objects and their co-occurrence information. The criterion of canonical *correspondence* analysis is similar to that of canonical *correlation* analysis,[26] so canonical correspondence analysis can be regarded as a variant of canonical correlation

analysis. The critical differences between them are as follows: i) the objects are the same across two different data in canonical *correlation* analysis, while the objects are different across two different data in canonical *correspondence* analysis, and ii) canonical correlation analysis cannot deal with co-occurrence information about the heterogeneous objects. Note that it is therefore impossible to directly apply the canonical correlation analysis or its sparse version[25,27,28] in the question addressed in this paper.

Evaluation of Extracted Components by Reconstruction of Drug-Target Interactions. We propose to evaluate the generalization properties of the method by the reconstruction of known drug-target interactions, based on the extracted substructures and protein domains.

Given a pair of compound $\mathbf{x}$ and protein $\mathbf{y}$, their potential interaction can be estimated based on the chemical substructures present in $\mathbf{x}$, the protein domains present in $\mathbf{y}$, their presence in common extracted components, and their distribution over all the canonical components. Therefore, we propose the following prediction score for any given pair of compound $\mathbf{x}$ and protein $\mathbf{y}$

$$s(\mathbf{x},\mathbf{y}) = \sum_{k=1}^{m} u_k \rho_k v_k = \sum_{k=1}^{m} \mathbf{x}^T \alpha_k \rho_k \beta_k^T \mathbf{y} \qquad (6)$$

where m is the number of canonical components, and ρ_k is the k-th singular value. If $s(\mathbf{x},\mathbf{y})$ is higher than a threshold, compound $\mathbf{x}$ and protein $\mathbf{y}$ are predicted to interact with each other.

We perform the following 5-fold cross-validation to evaluate the reconstruction ability. 1) We split drugs and target proteins in the gold standard set into five subsets of roughly equal sizes and take each subset in turn as a test set. 2) We perform the training of CCA model on the remaining four sets (i.e., we extracted canonical components based on the remaining four sets). 3) We compute the above prediction score for the test set, based on the components extracted from the training set. 4) Finally, we evaluate the prediction accuracy over the five folds.

Other Possible Methods To Reconstruct Drug-Target Interactions. If the CCA-based components capture relevant information about protein−ligand interactions, the performance in reconstruction of protein−ligand interactions should be high compared to that of other prediction methods. Therefore, we compare the performance of SCCA and OCCA to three other possible methods to retrieve drug-target interactions. Two are naive methods that can be considered as baseline methods, namely the Nearest Neighbor and Naive Bayes methods, and one is a state-of-the-art method, namely Pairewise Support Vector Machine. For all methods, we used the same representations for proteins and ligands, i.e. the feature vectors described in the Supporting Information.

Nearest Neighbor (NN). The classical nearest neighbor (NN) method is often used in molecular screening. The proteins potentially interacting with a newly given compound $\mathbf{x}$ are determined as those that interact with the most similar compound in the training set. The predicted interactions involving a new compound $\mathbf{x}$ are assigned prediction scores according to the fingerprint profile similarity between $\mathbf{x}$ and its nearest compound neighbor in the training data. Likewise, potential ligands for a newly given protein $\mathbf{y}$ are determined as those that bind to the most similar protein in the training set. The predicted interactions are assigned prediction scores according to the profile similarity between $\mathbf{y}$ and its nearest protein neighbor in the training data. The cosine correlation coefficient is used as a similarity measure for both compounds and proteins.

Table 1. Performance Evaluation on Drug-Target Interaction Reconstruction by 5-Fold Cross-Validation

	NN	Bayse	P-SVM	OCCA	SCCA
AUC	0.5892	0.5934	0.7504	0.7377	0.7497
SD.	0.0042	0.0062	0.0064	0.0046	0.0057

Naive Bayes (Bayes). The Bayesian classifier based on naive Bayes model is a binary classifier to predict the binary response based on high dimensional features and to select only a subset of features strongly correlated with the response. In our context, we test the naive Bayes method by concatenating drug substructure features and protein domain features into a single vector. We used the R package "predbayescor" which implements naive Bayes models with feature selection bias corrected.

Pairwise Support Vector Machine (P-SVM). The SVM is a well-known binary classifier, and it is becoming a popular classification method in bioinformatics and chemoinformatics because of its high-performance prediction ability.[29] The use of SVM with pairwise kernels have been proposed to predict new compound-protein interactions,[7−9] which is referred to as pairwise SVM (P-SVM). We tested several kernel functions such as linear kernel, Gaussian RBF kernel with various width parameters, polynomial kernel with various degree parameters for drug substructure profiles and protein domain profiles, and the regularization parameter.

RESULTS

Performance Evaluation for the Proposed Method. In general, it is difficult to evaluate the performance of a feature extraction method in a direct manner. However, if the extracted sets of chemical substructures and proteins domains (the components) are biologically meaningful and capture relevant information with respect to protein−ligand interactions, one would expect that they present good generalization properties. This can be evaluated by testing the ability of the method to reconstruct known drug-target interactions, using the prediction score and the 5-fold cross-validation scheme described in the Methods section.

We evaluated the performance of the method by the ROC (receiver operating characteristic) curve,[30] which 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. We summarized the performance by an AUC (area under the ROC curve) score, which is 1 for a perfect inference and 0.5 for a random inference. We repeated the cross-validation experiment five times and computed the average of the AUC scores over the five cross-validation folds, varying the three parameters of the method. The best results were obtained with $c_1 = 0.1$, $c_2 = 0.2$ for the sparcity parameters and with $m = 50$ for the number of components in the case of SCCA. The same experiments were repeated for OCCA which has only one parameter, and the best results were obtained for $m = 50$.

The AUC scores for SCCA and OCCA are 0.7497 and 0.7377, respectively. These statistics are summarized in Table 1. This result shows that both methods perform much better than a random inference, whose AUC score is equal to 0.5. Consequently, this indicates that the proposed prediction score allows to enlighten the good generalization properties of extracted SCCA or OCCA

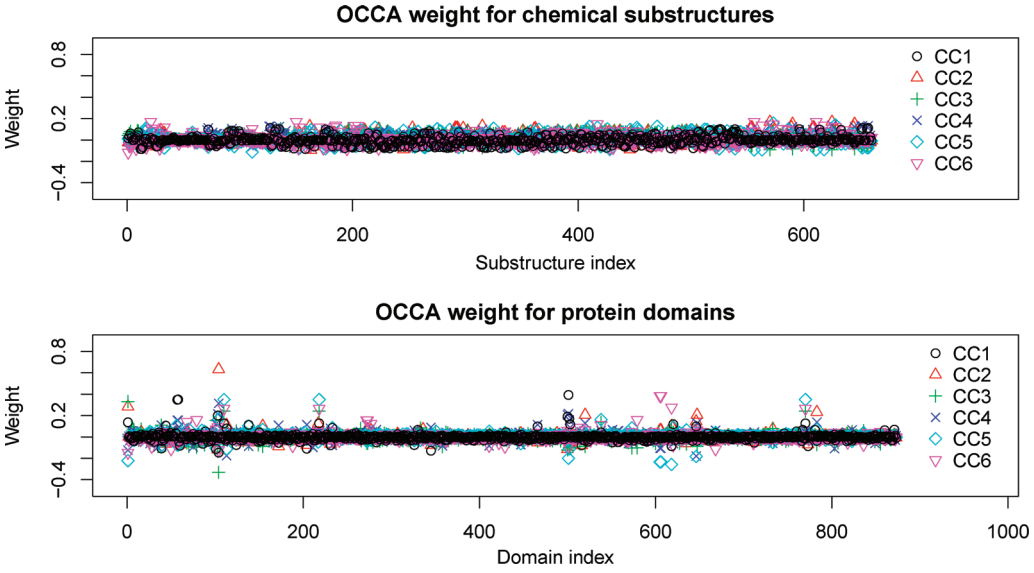


Figure 2. Index-plot of weight vectors for drug substructures and protein domains for OCCA. Horizontal axes indicate the index of chemical substructures (upper) or protein domains (bottom), and vertical axes indicate the weight values on the chemical substructures (upper) or protein domains (bottom).

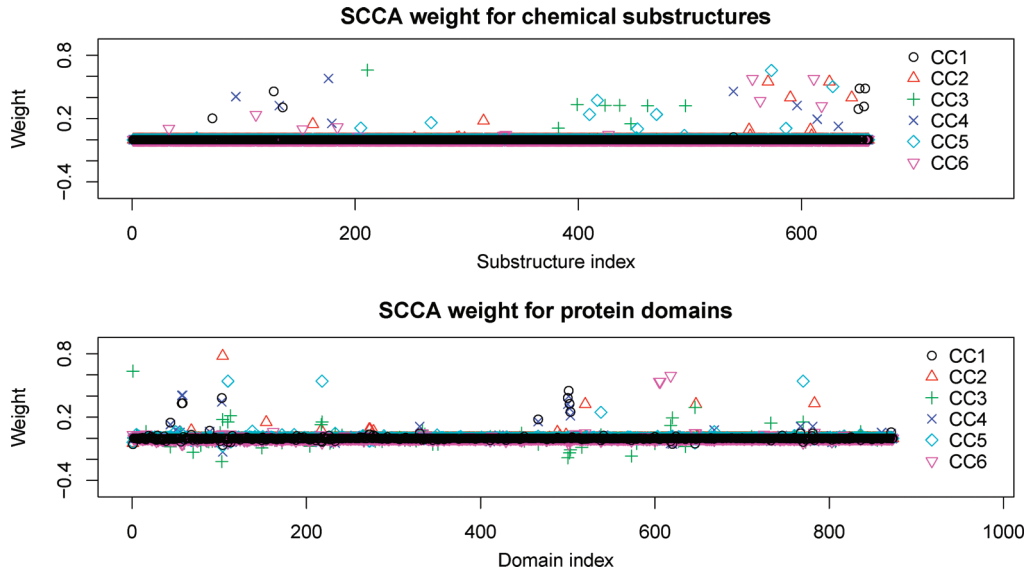


Figure 3. Index-plot of weight vectors for drug substructures and protein domains for SCCA. Horizontal axes indicate the index of chemical substructures (upper) or protein domains (bottom), and vertical axes indicate the weight values on the chemical substructures (upper) or protein domains (bottom).

components. Their performance comparison with the three other methods will be discussed in a later subsection.

Next, we applied SCCA and OCCA on the complete gold standard data set described in the Materials section and analyzed the extracted components of drug chemical substructures and protein domains. We used the parameters leading to the best results in the cross-validation experiment.

We examined the resulting weight vectors for drug chemical substructures and protein domains in applying OCCA and SCCA. Figure 2 shows the index-plot of weight vectors in applying OCCA, while Figure 3 shows the index-plot of weight vectors in applying SCCA, where the first six canonical components are shown in the both cases because of space limitation. It seems that almost all elements in the weight vectors in OCCA are nonzero and highly variable, while most of the elements in the weight vectors in SCCA are zero in each component, implying that SCCA can select a very small number of features as informative drug substructures and protein domains.

These results suggest that, although the performance in reconstruction of drug-target interactions of SCCA and OCCA were close, the proposed SCCA provides us with more selective drug substructures and protein domains, without missing important information encoding protein−ligand interaction. In practice, we found that it is very difficult to analyze the extracted components when there are too many high or low scoring weight elements like in OCCA. On the contrary, the advantage of SCCA over OCCA

Table 2. Examples of Results for Canonical Components 1, 2, 3, and 4: High Scoring Domains, Substructures, Proteins, and Drugs

CC1	protein domains	PF02159 (Estrogen receptor); PF02155 (Glucocorticoid receptor); PF00191 (Annexin);...
	drug substructures	CC1CC(O)CC1; CC1C(O)CCC1; saturated or aromatic carbon-only ring size 9; CC1C(C)CCC1;...
	proteins	ESR1_HUMAN (Estrogen receptor); GCR_HUMAN (Glucocorticoid receptor);...
	drugs	DB00443 (Betamethasone); DB00823 (Ethynodiol Diacetate); DB00663 (Flumethasone Pivalate));...
CC2	protein domains	PF00194 (Carbonic anhydrase); PF08403; PF02254; PF03493 (potassium channel);...
	drug substructures	SC1CC(S)CCC1; Sc1 cm^3(S)ccc1; Sc1c(Cl)cccc1; SC1C(Cl)CCCC1; N−S−C:C; N−S;...
	proteins	KCMA1_HUMAN (Calcium-activated potassium channel); CAH12_HUMAN (Carbonic anhydrase 12);...
	drugs	DB00562 (benzthiazide); DB00232 (Methyclothiazide); DB01324 (Polythiazide);...
CC3	protein domains	PPF00001 (transmembrane receptor); PF03491 (Serotonin neurotransmitter transporter);...
	drug substructures	C(H)(:C)(:C); C:C−C−C; C−C−C-C:C; C:C−C−C-C; C−C:C−C-C; C−C−C:C−C;...
	proteins	TOP2A_HUMAN (DNA topoisomerase); SC6A4_HUMAN (Sodium-dependent serotonin transporter);...
	drugs	DB01654 (Thiorphan); DB00743 (gadobenic acid); DB03788 (GC-24);...
CC4	protein domains	PF00105 (Zinc finger); PF00104; PF02159 (Estrogen receptor); PF00191 (Annexin);...
	drug substructures	C(C)(C)(C)(C); C−C(C)(C)−C-C; unsaturated nonaromatic carbon-only ring size 6;...
	proteins	ESR1_HUMAN (Estrogen receptor); GCR_HUMAN (Glucocorticoid receptor);...
	drugs	DB00596 (halobetasol); DB01234 (Dexamethasone); DB00620 (Triamcinolone);...

is that it is possible to derive biological interpretations, as shown on a few examples in the next subsection.

Extraction of Substructure-Domain Associations from Drug-Target Interactions. The SCCA method provided us with 50 canonical components. The output of the method is a list of canonical components (CCs), each of which contains correlated chemical substructures and protein domains, and a list of proteins and drugs that contributed to extract the chemical substructures and protein domains. All components present a limited number of high scoring chemical substructures and protein domains, which is a consequence of the sparsity of the method. It extracts domains and substructures that summarize the most relevant and consistent information. This allows meaningful analysis of the data for biological interpretation with respect to CCA.

Table 2 shows some examples of extracted chemical substructures (SMILE-like format in PubChem) and protein domains (PFAM IDs) in the first four CCs (CC1, CC2, CC3, and CC4). The results of all CCs can be obtained from Table S1 in the Supporting Information.

Since each component contains chemical substructures and protein domains which are associated with each other, the results of the method can be represented as a network. We evaluated the strength of the association between chemical substructures and protein domains by computing the product of their weight elements between chemical substructures and protein domains within each component. The merging of the results for all CCs provided us with a global view of substructure-domain association network estimated from the drug-target interaction network (for more details, see the Methods section), where we focused on chemical substructures and protein domains whose weights are higher than 0.1 for visualization simplicity. A graphical representation of this network for all components can be obtained from Figure S1 in the Supporting Information.

We also examined the chemical diversity of unique chemical substructures in the 50 components extracted by OCCA and SCCA. Figure 4 shows a scatter-plot of the number of unique chemical substructures that appear in at least one of the extracted components against the number of top ranked substructures on which we focused in each CC. All 660 substructures catch a positive weight in at least one component in the case of OCCA, while only 384 do in the case of SCCA. However, the same substructures appear many times in almost all different components in the case of OCCA, while the same substructures appear specifically in each component in the case of SCCA. This implies that OCCA components are more redundant than those in SCCA. In addition, we estimated the chemical diversity present in OCCA and SCCA components, by analyzing the number of unique chemical substructures appearing among the top ranked substructures. We observed that this number was larger for SCCA components than for OCCA components. For example, when considering the top 10 substructures, there are 264 unique substructures in OCCA, while 362 unique substructures in SCCA. This is an interesting observation, because in practical applications we can only use highly weighed substructures to derive a biological interpretation for each component. This result suggests that, even though OCCA takes into account the vast majority of the substructures, the most relevant substructures span a less diverse chemical space compared with SCCA.

Biological Interpretation of Extracted Drug Substructures and Protein Domains. We examined the extracted drug substructures and protein domains from biological viewpoints. The results for a few canonical components will be discussed because of the space limitation. Analysis of the results shows that the components contain a limited number high scoring protein domains that usually belong to one or a small number of protein families. For example, most high scoring protein domains of component CC1 belong to nuclear receptors (PF02159: Estrogen receptor, PF02155: Glucocorticoid receptor, PF00104: Ligand-binding domain of nuclear hormone receptor, PF02161: Progesterone receptor, PF02166: Androgen receptor, PF00105 zinc finger c4 type). Consistent with this observation, the high scoring substructures are typical fragments found in steroids, and the high scoring drugs are steroid-like molecules. The domains from nuclear receptors also appear with high scores in a few other components such as CC4 or CC12. However, these components do not share any of their high scoring chemical substructures, which shows that they are not redundant. We observed that the absence of redundancy between components is a general feature of the method.

Unexpectedly, the annexin domain PF00191 is also present in the top ranked domains of CC1. Annexins are membrane associated proteins that bind phospholipids, inhibit the activity of phospholipase A2, and play a role in the inflammatory response. Annexins and nuclear receptors are evolutionary unrelated proteins with no sequence or function similarities. However, annexins and nuclear

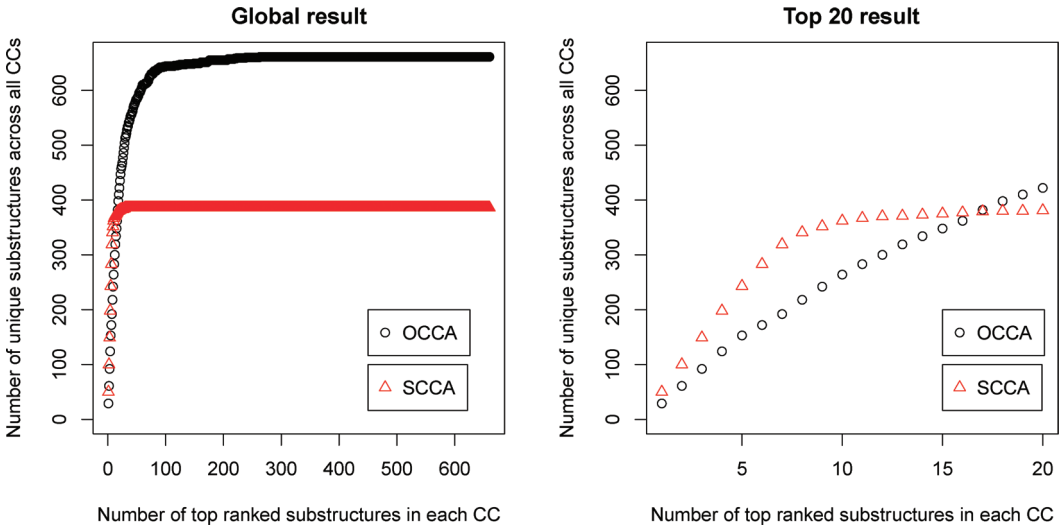


Figure 4. Chemical diversity of highly weighed substructures in OCCA and SCCA. Horizontal axes indicate the number of top highly weighted substructures on which we focused in each CC, and vertical axes indicate the number of unique substructures across all 50 CCs. Note that the right panel shows an image zooming version of the left panel where we focus on the top 20 substructures in the horizontal axis.

receptors probably present similar ligand binding pockets in the 3D space, which could not be foreseen from comparison of their primary sequences, and both types of proteins can bind steroid-like ligands. Therefore, the method associated these protein domains in CC1. Some of these steroids ligands are common to both types of proteins. For example DB00443 and DB00663 (respectively PubChem IDs 9782 and 443980) bind to glucocorticoid receptor and to annexin A1. On the contrary, some steroids only bind to a nuclear receptor and not to annexin. This observation suggests that the method might offer a tool to tackle the important question of specificity.

To illustrate this point, we will consider the example of the estrogen receptor ESR1_HUMAN (UniProt ID: P03372, Pfam IDs PF00104, PF00105, PF02159) and of annexin A1 ANXA1_HUMAN (UniProt ID: P04083, Pfam ID PF00191). Domains of these two proteins have high weights in a few common components (CC1, CC4, CC13, CC46, called group A), while only domains of estrogen receptors have high scores in components CC12, CC15, CC29, CC34, and CC38 (called group B), and only those of annexin have high scores in components CC19, CC20, CC21, CC26, and CC40 (called group C). We will show in the case of drug DB00823 (or PubChem ID 9270) that binds the estrogen receptor but not annexin, and of DB01013 (or PubCHem ID 32798) that binds to annexin but not to the estrogen receptor, how analysis of the substructures belonging to the A, B, and C groups can be used to explain the specificity of these two drugs. The parts of the chemical structure of DB00823 and DB01013 that can be built using high scoring substructures belonging to group A (components common to estrogen receptor and annexin) is colored in blue in Figure 5. They correspond to the main steroid scaffolds of these two molecules, as expected for proteins sharing similar types of ligands. However, additional chemical structures of the DB00823 molecule, colored in red in Figure 5, can only be built by using high scoring substructures found in components of group B, where only estrogen receptor domains have high scores. Similarly, additional chemical structures of DB01013, colored in red in Figure 5, can only be built using high scoring substructures found in components of group C, where only annexin domains have high scrores. Note that none of the high scoring substructures of components specific of estrogen receptor (group B) are present in DB01013 that only bind annexin and that reciprocally none of the high scoring substructures of components specific of annexin (group C) are present in DB00823 that only bind estrogen receptor. In other words, the method allows to highlight the parts of the molecules that encode for their specificity to bind only to estrogen receptor only or only to annexin.

One additional comment should be made: all known annexin ligands are steroids, while estrogen receptor domains also bind other types of molecules such as Tamoxifen (DB00675, PubChem ID 2733526) or other similar molecules such as Raloxifen. As shown in Figure 6, these molecules are very different from steroids. They lead to the CC34 and CC38 above-mentioned components, that have a high score for estrogen receptor domains and not for annexin domains because the latter do not bind these molecules. In Figure 6, the part of the Tamoxifen molecule that can be built using high scoring substructures from CC34 and CC38 is colored in blue.

Figure 7 shows a graphical representation of some extracted sets of drug chemical substructures and protein domains involving the components discussed above (e.g., CC1, CC4, CC12, CC21, CC26). This graph illustrates that some evolutionary unrelated proteins may display interaction profiles with chemical substructures that are closer than those of proteins belonging to the same family. For example, we see that Annexin domains and two ligand-binding domains of nuclear receptors (namely those of Glucocorticoid and nuclear hormone receptors) are all connected to some chemical substructures of the 5 components (CC1, CC4, CC12, CC21, and CC26). The interaction profiles of these two nuclear receptor domains is less similar to that of other nuclear receptors ligand-binding domains such as progesterone, androgen, and estrogen receptors which are connected to some chemical substructures of only 3 components (CC1, CC4, and CC12).

We will comment more briefly on component CC2, in order to show that the above observations also apply to other components and families of proteins. Component CC2 contains the carbonic anhydrase domain PF00194, which belongs to zinc metalloenzymes

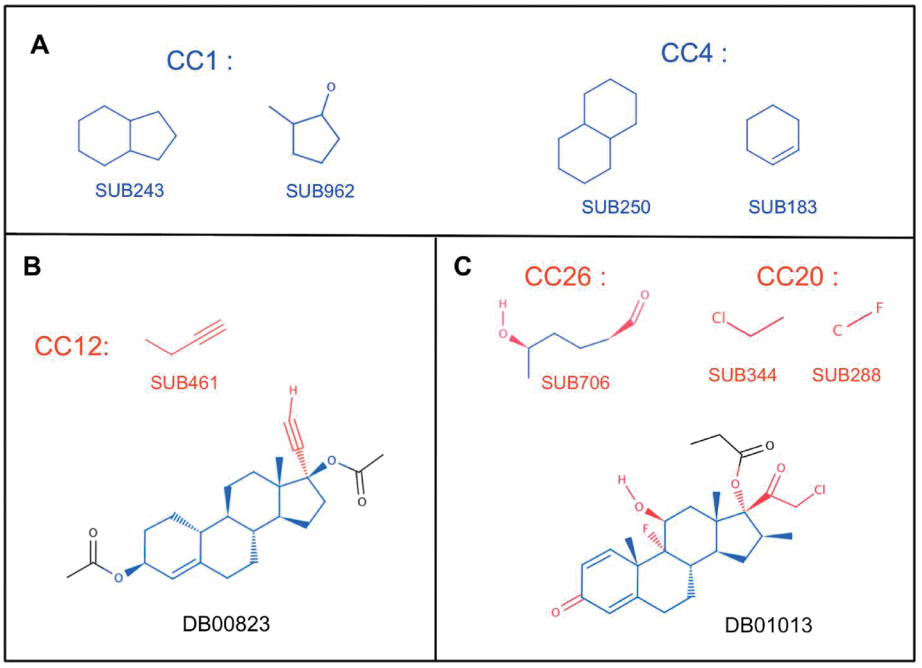


Figure 5. (A) Examples of high scoring substructures from components CC1 and CC4 belonging to group A. (B) In blue, part of the molecular structure of DB00823 that can be built using high scoring substructures of components of group A. In red, part of DB00823 that can be built using high scoring substructures of components of group B (specific of estrogen receptor). (C) In blue, part of the molecular structure of DB01013 that can be built using high scoring substructures of components of group A. In red, part of DB01013 that can be built using high scoring substructures of components of group C (specific of annexin). In the case of SUB706, only chemical groups that cannot be built using substructures of group A are colored in red.

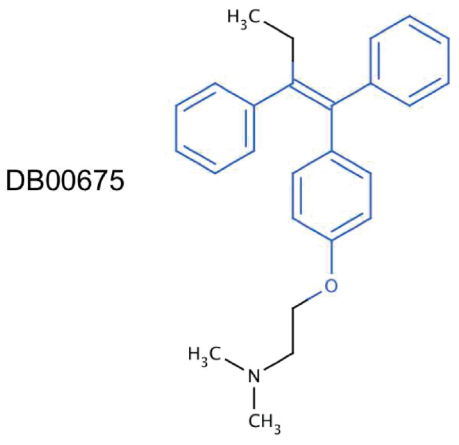


Figure 6. Chemical structure of Tamoxifen (DB00675). Parts of the molecule that can be built from chemical substructures of components CC34 and CC38 are colored in blue.

catalyzing reversible hydration of carbon dioxide to bicarbonate. It also contains calcium-dependent potassium channel domains (PF03493, PF02254). Carbonic anhydrase inhibitors are used as antiglaucoma agents, diuretics and antiepileptics. Interestingly, the human potassium channel KCMA1_HUMAN (UnitProt ID: Q12791), one of the high scoring proteins in CC2, is also known to be involved in epilepsy. Domains of the carbonic anhydrase and of the calcium-dependent potassium channel also appear together with high scores in a few other components (namely CC7, CC16, CC22, CC27, and CC43), whereas components CC3 and CC25 are specific of carbonic anhydrase, and component CC20 is specific of calcium-dependent potassium channel.

Although different types of drugs are known to bind human calcium-dependent potassium channel and carbonic anhydrase proteins, these two proteins share drugs from the thiazide family. Figure 8 shows the general scaffold of thiazide molecules. All known thiazide ligands of carbonic anhydrase also bind calcium-dependent potassium channel (for example DB00436 or DB00562, among others). However, the thiazide molecule DB00232 (PubChem ID 4121) only binds to KCMA1_HUMAN, the human calcium-dependent potassium channel, and not to human carbonic anhydrase. As in the case of annexins and nuclear receptors, although carbonic anhydrase and potassium channel present no sequence similarity, they share similar ligand binding pockets and are able to bind similar molecules. Therefore, the method associates them in CC2 and in a few other common components, namely CC7, CC16, CC22, CC27, and CC43. In Figure 8, the part of the DB00232 molecule that can be built using substructures of these components is colored in blue. However, the calcium-dependent potassium channel domains have high scores in component CC20, but this is not the case of carbonic anhydrase. In Figure 8, the part of DB00232 that can only be built using substructures of CC20 is colored in red. As in the case of estrogen receptor and annexin, the method allows to highlight the parts of the DB00232 that encode for its specificity to bind to calcium-dependent potassium channel and not to carbonic anhydrase.

Finally, we would like mention that component CC20 appears in the two cases discussed above, because the presence of SUB344 in steroid or thiazide molecules happens to modulate their specificity, respectively for annexin or calcium-dependent channel. The fact that a molecule contains substructure SUB344 does not necessarily mean that it will bind to all high scoring proteins of CC20. Indeed, more generally, the protein binding profile of a

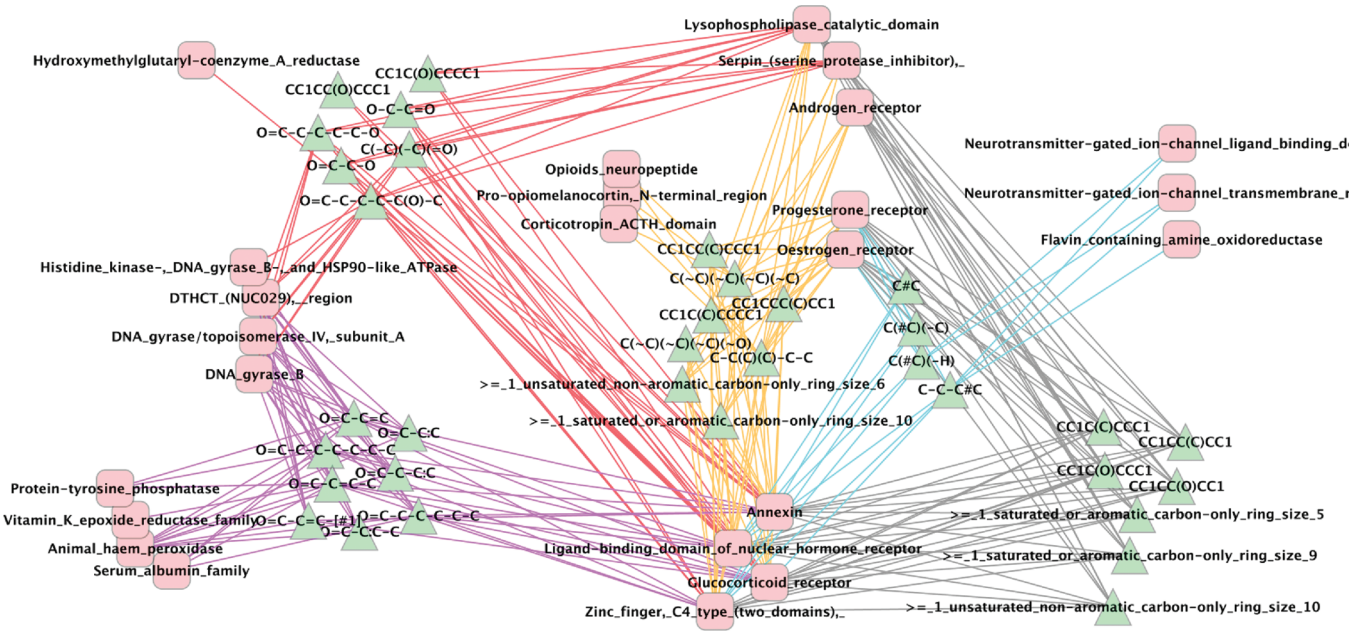


Figure 7. A graphical representation of chemical substructure-protein domain associations across different components. One green triangle corresponds to a chemical substructure, and one orange square corresponds to a protein domain. Edges are colored as follows: CC1 (gray), CC4 (yellow), CC12 (light blue), CC21(violet), and CC26 (red).

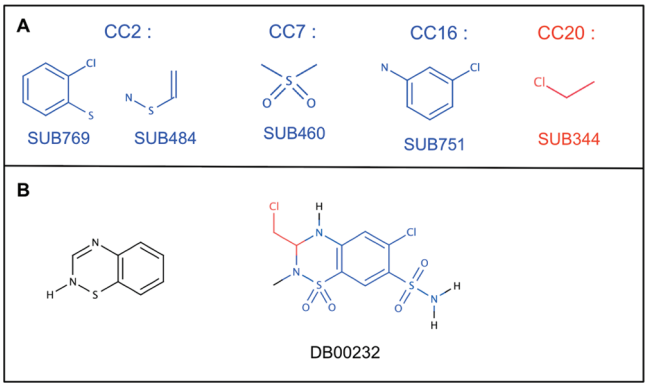


Figure 8. (A) In blue, examples of high scoring substructures of components common to calcium-dependent potassium channel and carbonic anhydrase proteins. In red, example of a high scoring substructure from component CC20 specific to calcium-dependent potassium channel. (B) On the left, in black, the basic thiazide scaffold. On the right, in blue, part of the molecular structure of DB00232 that can be built using high scoring substructures from components common to calcium-dependent potassium channel and carbonic anhydrase proteins. In red, part of DB00232 that can be built using high scoring substructures of components CC20, specific of calcium-dependent potassium channel.

molecule depends on its complete substructure profile which is not limited to an individual substructure.

Comparison with Other Related Methods. If the proposed method captures important features that govern protein−ligand interactions, and if the proposed prediction score is relevant, the performance should be at least as good as those of other methods for predicting protein−ligand interactions, using the same vector descriptions for proteins and ligands.

We performed the same 5-fold cross-validation experiments for the three other considered prediction methods NN, Bayes, and P-SVM on the same protein−ligand data set, as we did for SCCA and OCCA. The best performance was obtained for the number of retained features $k = 5$ in the case of Bayes, and for the polynomial kernel with degree parameter $d = 3$ and the regularization parameter $C = 1$ in the case of P-SVM .

Table 1 shows that SCCA and OCCA outperform the two methods used as baselines, i.e. the NN and Bayes. Furthermore, the performance of SCCA is similar to that of P-SVM used as the state-of-the-art prediction method. These results show that the extracted canonical components contain valuable biological information and underline the interest of the proposed method as a tool for analyzing protein−ligand interactions. They also show that the proposed prediction score (see eq 6 in section 3.3) is relevant, although this is not the main point of discussion here. In addition, it should be pointed out that P-SVM and NN do not provide any biological interpretation since they only predict interactions, and they do not extract any information about important molecular features for these interactions. The Bayes method enables us to extract discriminative features, but it is impossible to extract "correlated pairs" of substructure features and domain features as the SCCA does.

We also investigated the computational cost for each method. Figure S3 in the Supporting Information shows the total execution time of the cross-validation experiment between the five different methods in the scale of base10 logarithm. All methods were implemented in R on a Macbook(TM) with 2.16 GHz Intel Core 2 Duo processor and 2 GB RAM. It seems that NN is the fastest, followed by OCCA, SCCA, Bayes, and P-SVM. As expected, P-SVM is much slower than the other methods, because the complexity of the "learning" phase scales with the *square* of the "number of training compounds *times* the number of training proteins", leading to prohibitive computational difficulties for large-scale problems. These results suggest that the proposed method works the best in terms of prediction accuracy, biological interpretation, and computational efficiency.

DISCUSSION AND CONCLUSION

In this paper we proposed a novel method to extract drug chemical substructures and protein domains that govern drug-target interactions. The method uses all known protein−ligand interactions as a learning data set to extract ligand substructures and their associated protein domains, and importantly, it does not require information about proteins 3D structures. It provides an integrative analysis of chemical and genomic spaces in a unified framework, since the method can handle learning data sets containing many proteins from many different families. To our knowledge, there is no previous work reporting correlating chemical substructures to protein domains. Even if QSAR-like methods could in principle handle a single protein, it could not perform such correlation at a genome-wide level.

The method could be of interest in various ways in the drug development process. First, given protein target of therapeutic interest, one can identify the components in which this protein domains are found with high scores. Then, one can build a ligand for this target protein using high scoring substructures of these components, potentially with the help of other recent developments in fragment-based drug discovery.[31] For example, in the Results section, we showed for several drugs (DB00823, DB01013, DB00675, DB00232), that one could build their molecular structure using high scoring substructures of components in which their protein targets have a high score.

Second, for a given drug that binds to a protein target of interest, the method can help to identify off-target proteins. If some domains are found with high scores in the same components, the proteins with those domains are potential off-targets. Trivial off-targets are proteins that share high sequence similarity with the target protein and are otherwise easy to identify using classical algorithms such as BLAST.[32] However, two unrelated proteins that underwent convergent evolution may present similar pockets in the 3D space, allowing binding of similar ligands, although they may share no significant sequence similarity. The proposed method can handle such cases by learning ligand similarities from a database. Examples are shown in the Results section for estrogen receptor and annexin or for calcium-dependent potassium channel and carbonic anhydrase. Various methods have been developed to predict protein−ligand interactions, but predicting off-targets for a drug has been much less studied. As mentioned in the Introduction section, docking is a precious tool to predict or optimize ligands for a given protein. In principle, it could also be used to propose off-targets for a given drug, by docking the drug molecule against multiple targets. Potential off-targets would be proteins against which the drug has the best docking scores. A strong limitation is that docking require to know the protein 3D structure that is not always available. Moreover, even for proteins of known 3D structures, it is extremely difficult to automate the setup of these methods for many different binding sites, leading to docking and scoring inaccuracies when they are used on a large scale.[33] The use of drug side-effect similarity has been proposed to identify potential off-targets, but the method can therefore only be used for molecules of known side-effect profile, i.e. mainly for marketed drugs.[34] Prediction of all protein targets for a molecule is the goal of chemogenomics. However, only a limited number of studies report algorithms that implement such methods, and up to now, they have restrained the search for off-targets within a given family of proteins such as GPCRs.[35−37] The proposed method is a new contribution to this field. It does not require 3D structure proteins and only uses the chemical structure of the ligands, but it relies on learning databases of known protein−ligand interactions (DrugBank in this study, but it could be enriched by adding other known interactions found in other databases such as Kegg Ligand and Matador).

Finally, the method can also help to tackle the problem of drug specificity, which is in fact related to the topic of off-target identification. Here, we are interested in a given drug developed against a given target but that happens to also bind some off-target proteins. The method could help to optimize the structure of this drug. The principle would be to add chemical substructures that have high scores in components where only the target protein has a high score and not the off-target proteins. As shown in the Results section for estrogen receptor and annexin, drugs that bind only one of these proteins contain substructures present in components where only one of these two proteins has a high score. The same situation was shown for DB00232 that only bind calcium-dependent potassium channel and not carbonic anhydrase.

As shown in the Results section, the proposed method allows to identify such cases: nontrivial off-targets are expected to appear in the same components as the main target. Among several drug candidates, the method could help to eliminate molecules with too many potential off-target interactions or with potential off-target expected to lead to severe side effects. On the contrary, the drug candidates whose chemical substructures are not found in canonical components that do not contain its targeted protein domain are expected to be of greater interest.

From technical viewpoints, there are several limitations on the proposed SCCA method. One main difficulty of using SCCA is to choose appropriate sparsity parameters and appropriate number of components. High sparsity promoting parameters would lead to an oversparse model in all the cases, which might be misleading in the interpretation if the degree of sparsity was not tuned carefully. According to a cross-validation, we used the top 50 components, but other components may contain biologically meaningful information. The definition of an appropriate objective function to be maximized or minimized in the cross-validation is an important issue. There remains much room to develop a more appropriate way to choose the parameters. Another pitfall of SCCA is that it might not work well when sparsity is not a relevant characteristic arising from the data. For example, it cannot deal with hierarchically correlated features in the descriptors of drugs or proteins. Another possibility would be to additionally use other constraints which can deal with such a hierarchical effect.

In this study we applied our proposed method to known drugs, but any interactions between proteins and molecules are of interest, as soon as the corresponding protein and molecules can be represented using protein domains and molecular substructure descriptors. Additional interactions including nondrug molecules could significantly improve the diversity in the chemical data set. Moreover, the generalization properties of the model highly depend on the substructure descriptors, so the model cannot generalize to substructures absent from this learning set. The use of more complete descriptors such as Daylight and extented connectivity fingerprints[38] may improve the generalization properties of the model. A limitation of the method is that it uses substructure information of molecules, which encodes their 2D structures, while the protein−ligand recognition process takes place in 3D space. However, the active conformation of a drug, i.e. its conformation when bound to its protein target, is known for only a very limited number of drugs: those for which a crystal structure of the target-drug complex is available. Therefore, any other chemogenomics approach would face this limitation. Another

research direction is to extract informative chemical substructures and protein subsequences directly from the raw structured data (e.g., 2D or 3D graph structures for drugs, amino acid sequences for proteins) without using predefined feature representation for data objects. Recently, several structural data mining techniques have been proposed in order to extract complex features,[39] which do not require the predefinition of feature vectors representing each objects. A promising future work would be an extension of such mining techniques to a joint application of chemical graph mining and sequence mining for extracting drug substructures and protein domains.

■ ASSOCIATED CONTENT

Supporting Information. The whole result of extracted substructures and protein domains in all components (Table S1), a global substructure-domain association network for all components (Figure S1), and computational cost (Figure S3). This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: yoshihiro.yamanishi@ensmp.fr.

■ ACKNOWLEDGMENT

We thank Jean-Philippe Vert and Brice Hoffmann for useful discussions.

■ REFERENCES

(1) Kolb, P.; Ferreira, R.; Irwin, J.; Shoichet, B. Docking and chemoinformatic screens for new ligands and targets. *Curr. Opin. Biotech.* **2009**, *20* (4), 429–436.

(2) Kanehisa, M.; Goto, S.; Hattori, M.; Aoki-Kinoshita, K.; Itoh, M.; Kawashima, S.; Katayama, T.; Araki, M.; Hirakawa, M. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.* **2006**, *34*, D354–357.

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

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

(5) Gregori-Puigjané, E.; Mestres, J. A ligand-based approach to mining the chemogenomic space of drugs. *Comb. Chem. High. Throughput Screening* **2008**, *11* (8), 669–676.

(6) Keiser, M.;Predicting new molecular targets for known drugs. *Nature* **2009**, *462* (7270), 175–181.

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

(8) Faulon, J.; Misra, M.; Martin, S.; Sale, K.; Sapra, R. Genome scale enzyme-metabolite and drug-target interaction predictions using the signature molecular descriptor. *Bioinformatics* **2008**, *24*, 225–233.

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

(10) Yamanishi, Y.; Araki, M.; Gutteridge, A.; Honda, W.; Kanehisa, M. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* **2008**, *24*, i232–i240.

(11) Yamanishi, Y. Supervised bipartite graph inference. *Adv. Neural Inform. Process. Syst.* **2009**, *21*, 1841–1848.

(12) Klekota, J.; Roth, F. Chemical substructures that enrich for biological activity. *Bioinformatics* **2008**, *24* (21), 2518–2525.

(13) Han, L.; Wang, Y.; Bryant, S. Developing and validating predictive decision tree models from mining chemical structural fingerprints and high-throughput screening data in PubChem. *BMC Bioinf.* **2008**, *9*, 401.

(14) Shigemizu, D.; Araki, M.; Okuda, S.; Goto, S.; Kanehisa, M. Extraction and Analysis of Chemical Modification Patterns in Drug Development. *J. Chem. Inf. Model.* **2009**, *49*, 1122–1129.

(15) Morris, R.; Najmanovich, R.; Kahraman, A.; Thornton, J. Real spherical harmonic expansion coefficients as 3D shape descriptors for protein binding pocket and ligand comparisons. *Bioinformatics* **2005**, *21* (10), 2347–2355.

(16) Najmanovich, R.; Kurbatova, N.; Thornton, J. Detection of 3D atomic similarities and their use in the discrimination of small molecule protein-binding sites. *Bioinformatics* **2008**, *24* (16), i105–i111.

(17) Hoffmann, B.; Zaslavskiy, M.; Vert, J.; Stoven, V. A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. *BMC Bioinf.* **2010**, *22*, 99.

(18) Wishart, D.; Knox, C.; Guo, A.; Shrivastava, S.; Hassanali, M.; Stothard, P.; Chang, Z.; Woolsey, J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34*, D668–D672.

(19) Chen, B.; Wild, D.; Guha, R. PubChem as a Source of Polypharmacology. *J. Chem. Inf. Model.* **2009**, *49*, 2044–2055.

(20) The Uniprot Consortium. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* **2010**, *38*, D142–D148.

(21) Finn, R.; Tate, J.; Mistry, J.; Coggill, P.; Sammut, J.; Hotz, H.; Ceric, G.; Forslund, K.; Eddy, S.; Sonnhammer, E.; Bateman, A. The Pfam protein families database. *Nucleic Acids Res.* **2008**, *36*, D281–D288.

(22) Greenacre, M. *Theory and applications of correspondence analysis*; Academic Press: 1984.

(23) Dudoit, S.; Fridlyand, J.; Speed, T. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.* **2001**, 1151–1160.

(24) Tibshirani, R.; Hastie, T.; Narasimhan, B.; Chu, G. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statist. Sci.* **2003**, 104–117.

(25) Witten, D.; Tibshirani, R.; Hastie, T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **2009**, *10*, 515–534.

(26) Hotelling, H. Relation between two sets of variates. *Biometrika* **1936**, *28*, 322–277.

(27) Parkhomenko, E.; Tritchler, D.; Beyene, J. Genome-wide sparse canonical correlation of gene expression with genotypes. *BMC Proc.* **2007**, *1*, S119.

(28) Waaijenborg, S.; de Witt Hamer, P. V.; Zwinderman, A. Quantifying the Association between Gene Expressions and DNA-Markers by Penalized Canonical Correlation Analysis. *Stat. Appl. Genet. Mol. Biol.* **2008**, *7* (1), 3.

(29) Schölkopf, B.; Tsuda, K.; Vert, J. *Kernel Methods in Computational Biology*; MIT Press: 2004.

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

(31) Gozalbes, R.; Carbajo, R.; Pineda-Lucena, A. From fragment screening to potent binders: strategies for fragment-to-lead evolution. *Mini-Rev. Med. Chem.* **2009**, *9* (8), 956–961.

(32) Altschul, S.; Gish, W.; Miller, W.; Myers, E.; Lipman, D. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410.

(33) Kellenberger, E.; Foata, N.; Rognan, D. Ranking targets in structure-based virtual screening of three-dimensional protein libraries: methods and problems. *J. Chem. Inf. Model.* **2008**, *48*, 1014–1025.

(34) Campillos, M.; Kuhn, M.; Gavin, A.; Jensen, L.; Bork, P. Drug target identification using side-effect similarity. *Science* **2008**, *321* (5886), 263–266.

(35) Weill, N.; Rognan, D. Development and Validation of a Novel Protein- Ligand Fingerprint To Mine Chemogenomic Space: Application to G Protein-Coupled Receptors and Their Ligands. *J. Chem. Inf. Model.* **2009**, *49*, 1049–1062.

(36) Jacob, L.; Hoffmann, B.; Stoven, V.; Vert, J. Virtual screening of GPCRs: an in silico chemogenomics approach. *BMC Bioinf.* **2008**, *9*, 363.

(37) Yamanishi, Y.; Kotera, M.; Kanehisa, M.; Goto, S. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* **2010**, *26*, i246–i254.

(38) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754.

(39) Saigo, H.; Nowozin, S.; Kadowaki, T.; Kudo, T.; Tsuda, K.; gBoost, A Mathematical Programming Approach to Graph Classification and Regression. *Mach. Learn.* **2008**, *75*, 69–89.