February 2009

Dissertation for Doctor of Science

Genome-wide statistical prediction of interactions between biomolecules

Keio University
Graduate School of Science and Technology
School of Fundamental Science and Technology
Center for Biosciences and Informatics

Nobuyoshi Nagamine

Abstract

The fact that the genomes of more than 800 species have been completely decoded demonstrates that many data have been recently produced to elucidate the life. Along with the genome or the 'blueprint' of life, the molecular biology has accumulated the information of 'parts' constituting the life. However, even though with abundant information on 'blueprintI and 'parts', integration of these to design the life are still hard. Therefore, this study focused on prediction of interactions between biomolecules in attempt to contribute to overcoming these difficulties.

The biomolecules, including proteins and metabolites, constitute networks and systems to realize biological functions. Predicting interactions, which can be regarded as one of the minimal units of these systems, between them can contribute to the elucidation of biological mechanisms. In particular, the prediction of interaction between proteins relevant to diseases and small molecules can be of help in searching lead compounds in drug discovery and identifying unknown effects and side effects of known drugs, and can be of economic and industrial significance.

In Part 1 of this thesis, the significance of comprehensively predicting interactions between biomolecules in elucidating the biological system was explained.

In Part 2, the protein-protein interaction (PPI) network was utilized to identify cooperative elements in the transcription regulation network. In the computational experiment based on Chromatin Immuno-Precipitation (ChIP) and PPI data of yeast (S. cerevisiae), it was discovered that the transcription factors regulating proteins that were located close to one another in the PPI network tended to work cooperatively. This study also revealed meaningful relations between two different biological networks.

In Part 3, the prediction of interactions between proteins and chemical compounds was studied. In the computational experiment using the statistical learning method Support Vector Machine (SVM), relatively accurate prediction of interactions between approved drugs and their target proteins was achieved only by using easily available chemical structure and mass spectrometry data, and amino acid sequence data,. Moreover, in the comprehensive binding ligand prediction and the experimental verification with human androgen receptor, integration of these enabled more effective and efficient interaction predictions.

In Part 4, this study was summarized and future works were discussed.

Contents

Part I

Introduction

The fact that the genomes of 884 species have been completely decoded as of November 2008 (Liolios *et al.*, 2008) demonstrates that many data have been recently produced to elucidate the life. Along with the genome or the "blueprint" of life, the molecular biology has accumulated the information about "parts" constituting the life. On the basis of the accumulated information, there exist several attempts to design the life or understand the life as a system such as the synthetic biology (Andrianantoandro *et al.*, 2006; Benner and Sismour, 2005) and the systems biology (Kitano, 2002). However, even though with abundant information on "blueprint" and "parts", combination or integration of these to design the life are still hard particularly due to a large amount of possible combinations of biomolecules. For example, as the PubChem Compound database (`http://pubchem.ncbi.nlm.nih.gov/`) contains approximately 19 million compounds and UniProt (Apweiler *et al.*, 2004) includes 20,329 manually annotated human proteins as of November 2008, there exist 0.4 trillion possible pairs of proteins and chemical compounds. Therefore, this study focused on prediction of interactions between biomolecules in attempt to contribute to overcoming these difficulties.

The biomolecules, including proteins and metabolites, constitute networks and systems to realize biological functions. Predicting interactions, or one of the minimal units of these systems, between them can contribute to the elucidation of biological mechanisms. In particular, the prediction of interactions between proteins relevant to diseases and small molecules can be of help in searching lead compounds in drug discovery and identifying unknown effects and side effects of known drugs, and can be of economic and industrial significance (Gasteiger and Engel, 2003).

In order to deal with a huge amount of possible combinations, wide applicability and fast computation are essential. In order to realize these requirements, statistical methods are well used in bioinformatics and chemoinformatics (Bock and Gough, 2001; Karchin *et al.*, 2002; Zernov *et al.*, 2003; Bhasin and Raghava, 2004; Xue *et al.*, 2004; Bock and Gough, 2005; Jorissen and Gilson, 2005; Martin *et al.*, 2005; Swamidass *et al.*, 2005; Xiao *et al.*, 2006; Wang *et al.*, 2006; Yu *et al.*, 2006; Eckert and Bajorath, 2007; Hecht and Fogel, 2007; Jacob and Vert, 2008). Particularly, with respect to efficient utilization of accumulated data including BIND (an D. Betel and Hogue, 2003) and DIP (Xenarios *et al.*, 2002) for protein-protein interactions, and BindingDB (Liu *et al.*, 2007), BRENDA (Schomburg *et al.*, 2004), DrugBank (Wishart *et al.*, 2008), GLIDA (Okuno *et al.*, 2008), KEGG BRITE (Kanehisa *et al.*, 2006), PDSP Ki database (Roth *et al.*, 2000) and SuperTarget (Gunther *et al.*, 2008) for protein-ligand interactions, the statistical learning methods such as the support vector machines (Vapnik, 1998; Cristianini and Sawe-Taylor, 2000) and the artificial neural network (Ripley, 1996) have advantages to be applied.

In this study, the following two case studies were presented to show feasibility and effectivity of the comprehensive statistical prediction of interactions between biomolecules.

In Part II, the protein-protein interaction network was utilized to identify cooperative elements in the transcription regulation network. This approach is based on the assumption that proteins that are close to each other in the protein-protein interaction network, which tend to have common biological functions (Schwikowski *et al.*, 2000) and be involved in the same biological process, are likely to be co-regulated by a same set of transcription factors, and thus to be expressed at the same or similar timing.

In the computational experiment based on Chromatin Immuno-Precipitation (ChIP) (Lee *et al.*, 2002) and protein-protein interaction data (Mewes *et al.*, 2002; an D. Betel and Hogue, 2003; Xenarios *et al.*, 2002) of yeast (*S. cerevisiae*), it was discovered that the transcription factors that regulated proteins that were located close to one another in the PPI network tended to work cooperatively. This study also revealed meaningful relations between two different biological networks.

In Part III, the prediction of interactions between proteins and chemical compounds was studied. In the study, we attempted to realize wide applicability by using easily available data and constructing a multi-aspect model. The multi-aspect model was realized by considering pairs of proteins and chemical compounds, and can be applied to proteins that have no known binding ligands on the basis of the rules extracted from interactions between other proteins and their binding ligands. The multi-aspect model may contribute to improve the current situation of drug discovery in which most newly developed drugs have targeted proteins that is a target of previous drugs (Yildirim *et al.*, 2007) and there exist a number of druggable but undrugged proteins (394 drugged proteins (Yildirim *et al.*, 2007) vs. 3000 estimated druggable proteins (Russ and Lampel, 2005)).

In the computational experiment using the statistical learning method Support Vector Machine (SVM) (Vapnik, 1998; Cristianini and Sawe-Taylor, 2000) and based on easily available chemical structure and mass spectrometry data, and amino acid sequence data, relatively accurate prediction of interactions between the approved drugs and their target proteins was achieved. On the basis of this achievement, two types of comprehensive prediction approaches, comprehensive target protein prediction in which target proteins for a fixed chemical compound are predicted from databases of amino acid sequences such as UniProt (Apweiler *et al.*, 2004) and comprehensive binding ligand prediction in which binding ligands for a fixed protein are predicted from databases of chemical compounds such as PubChem Compound database, were conducted and both of these predictions produced biologically relevant results.

Moreover, in order to take advantage of use of SVM which can reflect feedback, we proposed an approach that effectively integrate computational predictions and experimental verifications. The approach was experimented in comprehensive binding ligand prediction of the human androgen receptor and the *in vitro* binding assay of

predictions to successfully identify novel ligands distant from known ligands in the chemical space.

Part II

Identifying cooperative transcriptional regulations using protein-protein interactions

1 Background

Promoter regions of higher eukaryotic genes have complex structures to regulate their transcriptional activations and are controlled by multiple transcription factors. Transcription factors (TFs, for short) are DNA-binding proteins at the terminals of signal transduction networks, and computational representations and identifications of binding sites for transcription factors have been widely studied (Stormo, 2000).

Recent molecular technologies have produced many kinds of experimental data including whole genome sequences, gene expression profiles, and protein-protein interactions. Therefore several methods have been developed to infer transcriptional regulation networks by combining these various kinds of data. For example, Banerjee and Zhang, 2003 used expression data and the chromatin immuno-precipitation (ChIP) data to predict cooperativity of transcription factors, which is a key factor for the analyses of complex transcriptional regulation networks. This line of approach is very significant because according to the fact that there are at most 200 different TFs among totally 6,300 genes in yeast organism, there must exist cooperative transcriptional activations to control the expressions of all 6,300 genes.

Recently, in addition to gene expression data, protein-protein interaction data have been rapidly generated. In this study, we propose a method integrating protein-protein interaction data and ChIP data (our strategy is illustrated in Fig. II.1), and show the effectiveness of exploiting protein-protein interactions to identify cooperative transcriptional activations. The existence of interaction between two proteins suggests that they contribute to the same or similar biological processes. Many cellular processes and chemical events in organisms such as enzymatic reactions and dimerization involve protein-protein interactions. In addition, these interactions reveal, in some cases, functional similarity of proteins. Schwikowski *et al.*, 2000 proposed a protein function prediction method by which a protein of unknown function is predicted to have the three most frequent cellular functions represented among its direct interaction partners.

On the other hand, gene expression data have also been applied to protein function predictions. Clustering analysis of gene expression data can be used to predict functions of unannotated proteins based on the idea that genes with similar functions are likely to be co-expressed (Eisen *et al.*, 1998; Brown *et al.*, 2000).

Based on these observations, we may deduce that the existence of protein-protein interaction is strongly related to the correlation of gene expressions from the viewpoint of functional similarity. This leads to the fundamental assumption of this study that proteins that are close to each other in the protein-protein interaction network

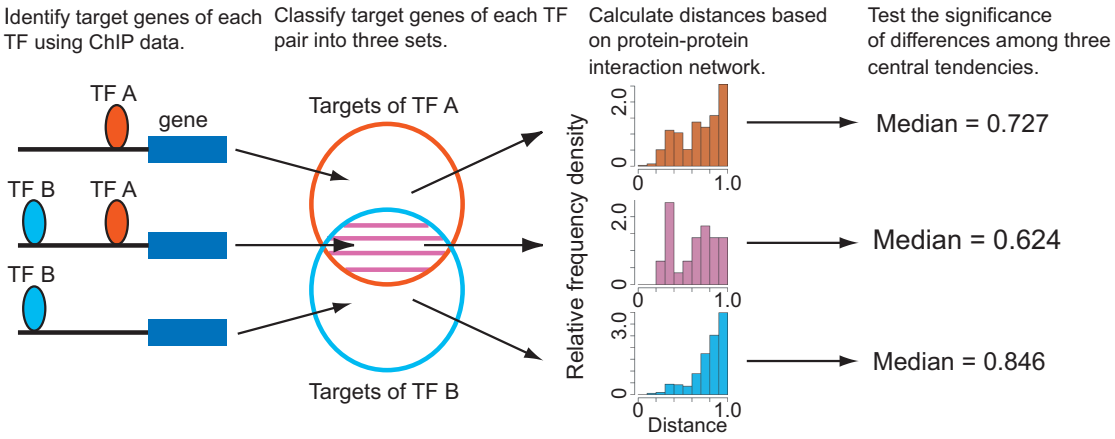


Fig. II.1 Strategy of identifying cooperative TF regulation using ChIP data and protein-protein interaction data. If the central tendency (we use median) of the overlap set that includes genes which both TF A and TF B bind to is significantly lower than those of the other two sets, we conclude that TF A and TF B are cooperative.

are likely to be co-regulated by a same set of transcription factors. In fact, this assumption is supported by the observation made by (Jansen *et al.*, 2002) that protein complexes have a strong relationship with gene expressions. Therefore, we can use the similarities of biological processes measured by protein-protein interactions to identify the cooperative transcription factors.

In our method, first, the protein-protein interaction networks are used to calculate the similarity of biological processes that the genes contribute to. Second, we integrate the similarity of biological processes based on protein-protein interactions and ChIP data to identify synergistic binding of transcription factors. Our computational experiments in yeast show that predictions made by our method based on protein-protein interactions have successfully identified eight pairs of cooperative transcription factors that have literature evidences and could not be identified by the previous method based on gene expression data. In addition, twelve new possible pairs of cooperative transcription factors have been inferred. From our careful analyses of the biological relevances for those pairs, we suggest a biological observation that some metabolism is regulated rather on translation level than on transcription level.

Furthermore, when using protein-protein interaction data, a typical problem is their noisiness, that is, the data contains many false-positives. Integrating various kinds of data is one solution to this problem (Jansen *et al.*, 2003). In addition, it may enable us to take many biological perspectives into consideration. In this study, we propose a method integrating cellular localization data and function data with protein-protein interaction data for precise predictions of TF cooperativity.

2 Methods

2.1 Selecting target genes of TF pairs from ChIP data

We use Lee *et al.*, 2002's ChIP data, a genome-wide binding data of 113 yeast regulators, to determine target genes of each transcription factor.

We suppose that a protein is regulated by a TF if its binding P-value $P_B < 0.001$ is satisfied. For all pairs of TFs A and B, we divide target genes for two TFs into three sets: those of TF A but not B (that is, $A \cap \bar{B}$), TF B but not A ($\bar{A} \cap B$), and both TF A and B ($A \cap B$). TF pairs whose overlap set ($A \cap B$) has genes less than threshold $O_{\min}$ are excluded.

2.2 Constructing protein-protein interaction network

We construct a protein-protein interaction network based on the dataset organized by Yu *et al.*, 2004. It contains data produced by different experiments (compiled from MIPS (Mewes *et al.*, 2002), BIND (an D. Betel and Hogue, 2003) and DIP (Xenarios *et al.*, 2002) databases), those by large-scale yeast two-hybrid experiments and those by *in vivo* pull-down experiments. It consists of 69,592 interactions involving 4,957 proteins. The average number of interaction partners per protein is about 27.9.

Fig. II.2 shows the number of interactions in each dataset. The distribution of D_G, the minimum number of edges needed to traverse from one protein to the other, in the dataset is shown in Fig. II.3.

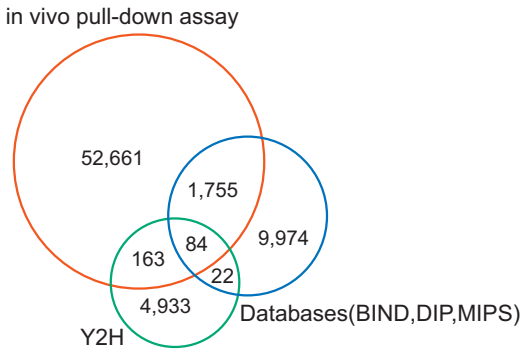


Fig. II.2 The distribution of interactions.

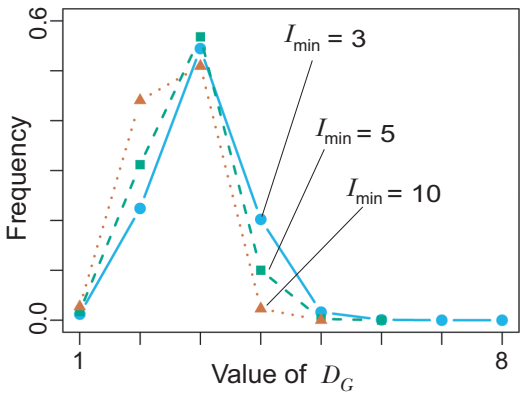


Fig. II.3 The distribution of D_G. When $I_{\min}$, which determines proteins whose distances are calculated, equals to 1 or 2, some distances between proteins can't be defined and thus the distribution on these $I_{\min}$ are omitted.

2.3 Calculating distance between two proteins based on protein-protein interaction

We calculate a distance between any two proteins based on a newly defined distance function exploiting the protein-protein interaction network constructed as above.

Two typical distance measures between two proteins based on the protein-protein interaction network are a graph-theoretic distance D_G, and the Czekanowski-Dice distance D_{CD} proposed in Brun *et al.*, 2004.

For the protein i and the protein j, $D_G(i,j)$ is defined as the minimum number of edges needed to traverse from i to j. On the other hand, $D_{CD}(i,j)$ is defined as follows:

$$D_{CD}(i,j) = \frac{|Int(i)| + |Int(j)| - 2\,|Int(i) \cap Int(j)|}{|Int(i)| + |Int(j)|} \tag{II.1}$$

where $Int(i)$ and $Int(j)$ are the lists of interactors of the protein i and protein j plus themselves (to decrease the distance between proteins interacting with each other). D_{CD} ranges from 0 to 1. A feature of these distances is that the less the distance between two proteins is, the stronger their biological or functional relatedness is thought to be.

However, a serious problem of these distances is that they can't express diversity and specificity of distances between proteins adequately.

D_G is a discrete measure and cannot be defined for proteins that are not linked in the network. D_{CD} is less than 1 only if two proteins are within a distance of 2 in terms of D_G (otherwise, $|Int(i) \cap Int(j)|$ in equation (II.1) always equals to 0), while $D_G > 2$ for most pair of proteins in the gene sets of Yu *et al.*, 2004.

To overcome this problem, we extend the Czekanowski-Dice distance as follows:

$$D_I(i,j,l) = \frac{\displaystyle\sum_{k=1}^{l} \frac{1}{k}(|Int_k(i)| + |Int_k(j)|) \;-\; 2\sum_{n=1}^{l}\sum_{m=1}^{l} \frac{2}{m+n}|Int_m(i) \cap Int_n(j)|}{\displaystyle\sum_{k=1}^{l} \frac{1}{k}(|Int_k(i)| + |Int_k(j)|)} \tag{II.2}$$

where $Int_k(i)$ is a list of proteins whose D_G from the protein i is equal to k ($Int_1(i)$ includes the protein i itself as in the Czekanowski-Dice distance), and l denotes the range of D_G to be considered. $D_I(i,j,1)$ is equal to D_{CD}.

Finally, we define the new distance function between the protein i and j, denoted $D(i,j,l)$, as

$$D(i,j,l) = \min_k D_I(i,j,k), \quad k \le l \tag{II.3}$$

We use $D(i,j,2)$ as a protein distance and represent it as $D(i,j)$.

Further, we consider the protein pair as biologically significant only when both proteins have direct interactors more than threshold $I_{\min}$ (we use 3 as $I_{\min}$).

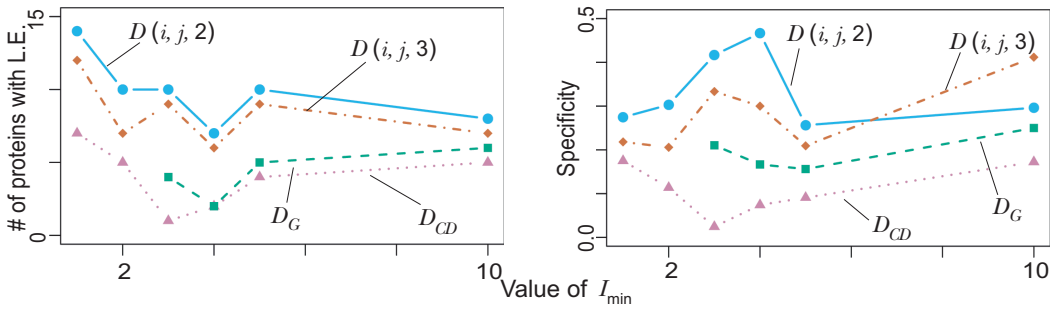


Fig. II.4 Comparison by other distance functions. $P_B \leq 0.001$, $O_{\min} = 3$ and $P_{mw} \leq 0.05$ are used. The number of predictions with literature evidence is relevant to sensitivity of each condition. Specificity equals to (# of predictions with literature evidence)/(# of predictions). As for D_G and D_{CD}, mean is used instead of median as the central tendency. There are proteins whose distances are not defined in using D_G with $I_{\min} = 1$ or 2. Thus, results on D_G with $I_{\min} = 1$ and 2 are omitted.

2.4 Evaluating transcription factor cooperativity

We assume that proteins which are close to each other in the protein-protein interaction network, or those which may contribute to the same biological processes, are regulated under the same control mechanism. On this assumption, if TF A and TF B are cooperative, proteins that are controlled by both TFs must be closer to each other in terms of $D(i, j)$ than those regulated only by TF A or TF B. To precisely measure this differences of distance among three gene sets, we examine whether the central tendency of the overlap distance set, which is a set of distances $D(i, j)$ for pairs of proteins i, j in the overlap gene set $(A \cap B)$, must be significantly lower than those for the two other distance sets, TF A distance set for the TF A gene set $(A \cap \bar{B})$ and TF B distance set for the TF B gene set $(\bar{A} \cap B)$.

First, we choose TF pairs in which the median of overlap distance set was lower than the other distance sets. We determine the significance of differences by the Mann-Whitney U test. As the distribution of distances is not normal (Fig. II.4), we use this non-parametric statistical test.

For a pair of TF A and TF B, if the P-value of Mann-Whitney U test for the combination of the overlap distance set and TF A distance set and that for the combination of the overlap distance set and the TF B distance set satisfy the threshold of 0.05 with Holm's correction, which means that the lower of these two P-values must be under 0.025 and the other P-value must be under 0.05, we conclude that TF A and TF B are cooperative.

The P-value for the Mann-Whitney U test is calculated as follows (Zar, 1998). First, the statistic U is calculated,

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - \sum_i r_{1i},$$

Fig. II.5 Interaction by localization algorithm

Input	protein i and protein$j$: Protein interaction data suggest their interaction.
Output	*True* or *False*: Whether their interaction is supported by localization data.
1)	Interaction(i,j)=*False*
2)	for l_i in L_i do
3)	for l_j in L_j do
4)	if $(I_m(l_i, l_j) > I_{\mathrm{loc}})$ Interaction(i,j)=*True*

where n_1 denotes the number of elements in 1st set, n_2 that of 2nd set,

and r_{1i} is a rank of i-th element in 1st set where ranks are assigned according to all the elements. Second, the statistic Z_0 is calculated,

$$Z_0 = \frac{|U - n_1 n_2/2|}{\sqrt{\frac{n_1 n_2}{12(n^2-n)}\left\{n^3 - n - \sum_{i=1}^{m}(t_i^3 - t_i)\right\}}},$$

where n denotes the number of all the elements, m the number of different kinds of ranks and t_i the number of elements of i-th rank.

Finally, the P-value of Z_0 is calculated based on the normal distribution. We execute one-tailed testing so that we only examine that the central tendency of the overlap set is significantly lower than those of the other two sets.

2.5 Prediction of cooperative TF triads

When the pair of TF A and TF B and that of TF B and TF C are both cooperative, we apply the same method to the triad of TF A, TF B and TF C.

2.6 Prediction of cooperative TF modules

When we consider TF A and TF B, there could be a case that the target genes of each TF are also those of other (third) TF C and TF D which are possibly cooperative. If the targets of cooperative transcription factors TF C and TF D are included in the TF A set, proteins in the TF A set can be quite close to each other in terms of $D(i, j)$ due to the influences of TF C and D. In that case, they may obscure the differences between the TF A distance set and the overlap distance set , and so obstruct the detection of their cooperativity.

To treat this problem, in judging cooperativity of TF A and TF B as above, we redefine the TF A gene set as genes to which only TF A binds, the TF B gene set in the same way, and the overlap gene set as those that both TF A and TF B, but no other TFs bind to. We apply this method to any combination of TFs whose TF and overlap sets have more than $O_{\min}$ genes. We represent combinations of TFs that are predicted as cooperative by this method as cooperative TF "modules".

2.7 Integrating other kinds of biological data

2.7.1 Cellular localization data

In high-throughput data of protein-protein interactions, many unnatural data such as an interaction between a protein in the nucleus and one in the plasma membrane are included. However, in the living cell, these interactions between proteins existing apart could be never observed. Therefore, we incorporate localization data of proteins to exclude unnatural interactions, and improve the reliability of protein-protein interactions.

A straightforward introduction of localization information is that for two proteins which have an interaction reported in experiments, if these two proteins have at least one common localization, then we accept the interaction, otherwise reject the interaction. However, since not all localization data for the proteins are available, this straightforward method is naive. We use mutual information criterion on co-occurrence of localizations in proteins to judge the relatedness of two localizations, and verify the possibility of protein-protein interactions. Mutual information I_m is expressed as follows:

$$I_m(i,j) = \sum_{k=\{0,1\}} \sum_{l=\{0,1\}} P(X_i = k, X_j = l) \log_2 \frac{P(X_i = k, X_j = l)}{P(X_i = k)P(X_j = l)}, \tag{II.4}$$

where i and j denote two localizations, $P(X_i = 1)$ denotes the probability that a protein has localization i, $P(X_i = 0)$ the probability that a protein doesn't have localization i, $P(X_i = 1, X_j = 1)$ the probability that a protein has both localization i and j. The mutual information is a method often used to give a quantitative relation between two discrete elements, and produces biologically relevant results applied to biological data (Huynen *et al.*, 2000; Sprinzak and Margalit, 2001).

For two proteins i and j which have an interaction reported in experiments, if there exist at least one combination of localization l_i that i has and l_j for j in which $I_m(l_i, l_j)$ exceeds some threshold I_{loc}, then we adopt the interaction between i and j (Fig. II.5).

We use MIPS cellular localization data (Mewes *et al.*, 2002) to calculate mutual information of localizations where the frequency of localization in the MIPS cellular localization data is regarded as the probability with respect to localization in Eq. (II.4) and to select reliable interactions.

2.7.2 Function data

The existence of protein-protein interaction often reflect functional similarity. Thus, we may incorporate function data annotated for proteins to refine the interaction-based protein distances. Here, we use the logistic regression to achieve this purpose. The logistic regression is a method, often used in medicine (Marchand *et al.*, 2001), to quantitatively evaluate a relation between one discrete element and the others. As we here want to know not a relation between two functions but that between a protein which may have multiple functions and a function, the logistic regression

Fig. II.6 Functional distance algorithm

Input	TF A, TF B, protein i, protein j
Output	$D_f(i,j)$: Functional distance between protein i and protein j
	F_{TF}: List of functions that TF A and TF B have in common
	F_t: List of functions that i and j have in common, and not included in F_{TF}
1)	for f_A in F_A do
2)	for f_B in F_B do
3)	if($I_m(f_A, f_B) > \alpha = 0.01$) add f_A and f_B to F_{TF}
4)	for f_i in F_i do
5)	for f_j in F_j do
6)	if($I_m(f_i, f_j) > I_{\text{fnc}}$))
7)	add f_i, f_j to F_t
8)	for f_{TF} in F_{TF} do
9)	if($I_m(f_i, f_{TF}) > I_{\text{fnc}}$) remove f_i from F_t
10)	if($I_m(f_j, f_{TF}) > I_{\text{fnc}}$) remove f_j from F_t
11)	$max = \max_{f_t \in F_t} \{\min(D_{CD_f}(i,j,f_t), D_{I_f}(i,j,2,f_t))\}$
12)	$D_f = \max\{\min(D_{CD}(i,j), D_I(i,j,2)), max\}$

is more appropriate than the mutual information. In addition. its output is more meaningful as probability and more tractable with the range [0, 1] than the linear multiple regression or the mutual information.

By using the logistic regression on co-occurrence of functions in proteins, we can calculate the possibility that a protein has some function z as follows:

$$Pr(z = 1|\boldsymbol{x}) = \frac{1}{1 + \exp\{-(1, \boldsymbol{x})^t \boldsymbol{\beta}\}} \tag{II.5}$$

where a vector $\boldsymbol{x}$ denotes the list of functions that a protein has excluding z, in which $x_i = 1$ if a protein has the function i and 0 if not, and a vector $\boldsymbol{\beta}$ consists of intercept and coefficients for each element in $\boldsymbol{x}$ and is estimated based on logistic regression model.

For a protein x and a function f, using the equation (II.5), we denote $P(x, f)$ as follows:

$$P(x,f) = \begin{cases} 1 & \text{if } F_f(x) = 1 \\ Pr(F_f(x) = 1|\boldsymbol{F(x)}) & \text{otherwise} \end{cases} \tag{II.6}$$

where $\boldsymbol{F(x)}$ denotes a functional vector of x.

Exploiting $P(x, f)$, we redefine the equations (II.2) and (II.3) for protein distances as follows:

$$\begin{aligned} f(Int_k(i)) &= \sum_{x \in Int_k(i)} P(x,f) \\ D_{If}(i,j,l,f) &= \frac{\sum_{k=1}^{l} \frac{1}{k}\{f(Int_k(i)) + f(Int_k(j))\} \; - \; 2\sum_{n=1}^{l}\sum_{m=1}^{l} \frac{2}{m+n} f(Int_m(i) \cap Int_n(j))}{\sum_{k=1}^{l} \frac{1}{k}\{f(Int_k(i)) + f(Int_k(j))\}} \\ D_f(i,j,l,f) &= \min_k D_{If}(i,j,k,f), \quad k \le l \end{aligned}$$

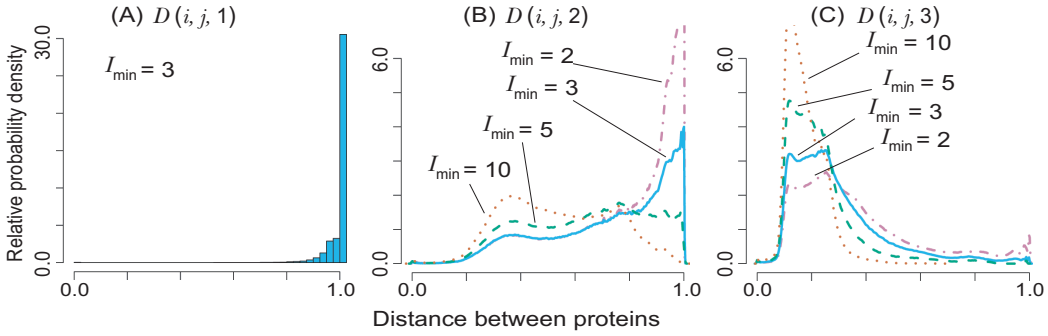


Fig. II.7 Distributions of distances based on the protein-protein interaction network $D(i, j, l)$. For all possible pairs of proteins that have more than $I_{\min}$ interactors, $D(i, j, l)$ is calculated. As for $D(i, j, 2)$ and $D(i, j, 3)$, the probability density on several $I_{\min}$ is estimated using Kernel method. $D(i, j, l)$ ranges from 0.0 to 1.0, and its distribution depends on l, the value determining the extent of interactors of protein i (j) to be considered.

We represent $D_f(i, j, 2, f)$ as $D_f(i, j, f)$.

If two proteins i and j have some common functions that are not related to the functions of the objective TFs, then we calculate $D_f(i, j, f)$ by choosing, among functions which two target proteins have in common, a function f which maximizes the distance as the proteins may be strongly influenced by other TFs (Fig. II.6).

We use the mutual information to measure the "relatedness" of protein functions with some threshold I_{fnc} as we have done in localization data.

We use MIPS function data (Mewes *et al.*, 2002) to construct logistic regression model and calculate functional distance D_f. We exclude proteins whose function is unknown when constructing logistic regression model.

2.8 Comparison with a method using expression data

We compare prediction results of our method with those of Banerjee and Zhang, 2003. The method of Banerjee *it al.* calculates the correlations of expression profiles for the target genes and finds cooperative TFs based on the correlations. Therefore, in order to investigate the relationships between predictions on which two methods agree or disagree, we calculate the relationships between the protein distances by our method based on protein-protein interactions and the correlations of gene expression profiles. We exploit genome-wide cell cycle expression data (Cho *et al.*, 1998) to obtain the correlation of expression profiles.

3 Results

3.1 Interaction based protein distance

First, we verify the adequacy of using extended Czekanowski-Dice distance $D(i, j, l)$ as a distance measure.

Fig. II.7 shows a distribution of $D(i, j, l)$. As shown in the figure, the distribution mainly depends on l, which determines the extent of interactors to be considered. The bigger l is, the more distant interactors are taken into consideration. $D(i, j, 1)$, or $D_{CD}(i, j)$, can't express variety of protein distances properly in which most distance is equal to 1. On the other hand, in using $D(i, j, 3)$, specificity of closeness may be lost as randomly chosen two proteins tend to be rather close to each other. Thus, it is reasonable to use $D(i, j, 2)$ as a protein distance.

3.2 Predictions of synergistic binding

From ChIP data of Lee *et al.*, 2002, with the threshold of binding P-value $P_B < 0.001$, which determined target genes of each TF, we extracted 476 TF pairs satisfying $O_{\min} = 3$, the threshold for the number of genes in overlap sets. TF pairs whose overlap set has more than $O_{\min}$ genes are considered. From these TF pairs, by calculating $D(i, j, 2)$ for pairs of proteins i and j both of which had more than the threshold $I_{\min} = 3$ direct interactors and judging the significance of Mann-Whitney tests on distance sets with $P_{mw} \leq 0.05$, we identified 24 pairs as cooperative TFs (Table II.1). Fig. II.4 shows the results of predictions on several parameter values for calculation of protein distances in terms of specificity and sensitivity based on the literature (see supplementary materials for details of predictions). It explains the reason to choose these parameters ($P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 3$), and indicates that parameter values which we chose produce the best result.

Compared with the literature and the predictions using gene expression data (Banerjee and Zhang, 2003), about half of our predictions overlap with their results (Table II.1, and Fig. II.8A, C).

3.2.1 Overlaps with literature

It is remarkable that cooperative TF pairs Hap2/Hap5, Hap3/Hap5, Arg80/Arg81, Fkh1/Fkh2, Fkh1/Ndd1, Mbp1/Skn7, Gcr1/Gcr2 and Skn7/Yap1 are only detected by our method and have literature evidences. Particularly, the detection of Hap2/Hap3/Hap5 cooperativity (Table II.2) meets with the fact that these three TFs form a hetero-trimer to be a CCAAT DNA-binding factor (McNabb *et al.*, 1995).

A half of the overlaps, including Hir1/Hir2, Fkh1/Fkh2, Fkh1/Ndd1, Fkh2/Mcm1 and Mbp1/Skn7, are involved in the cell cycle. Fkh1, Fkh2, Mcm1 and Ndd1 are TFs that mainly control the S/G_2, G_2 and G_2/M phases (Simon *et al.*, 2001). Mbp1 and

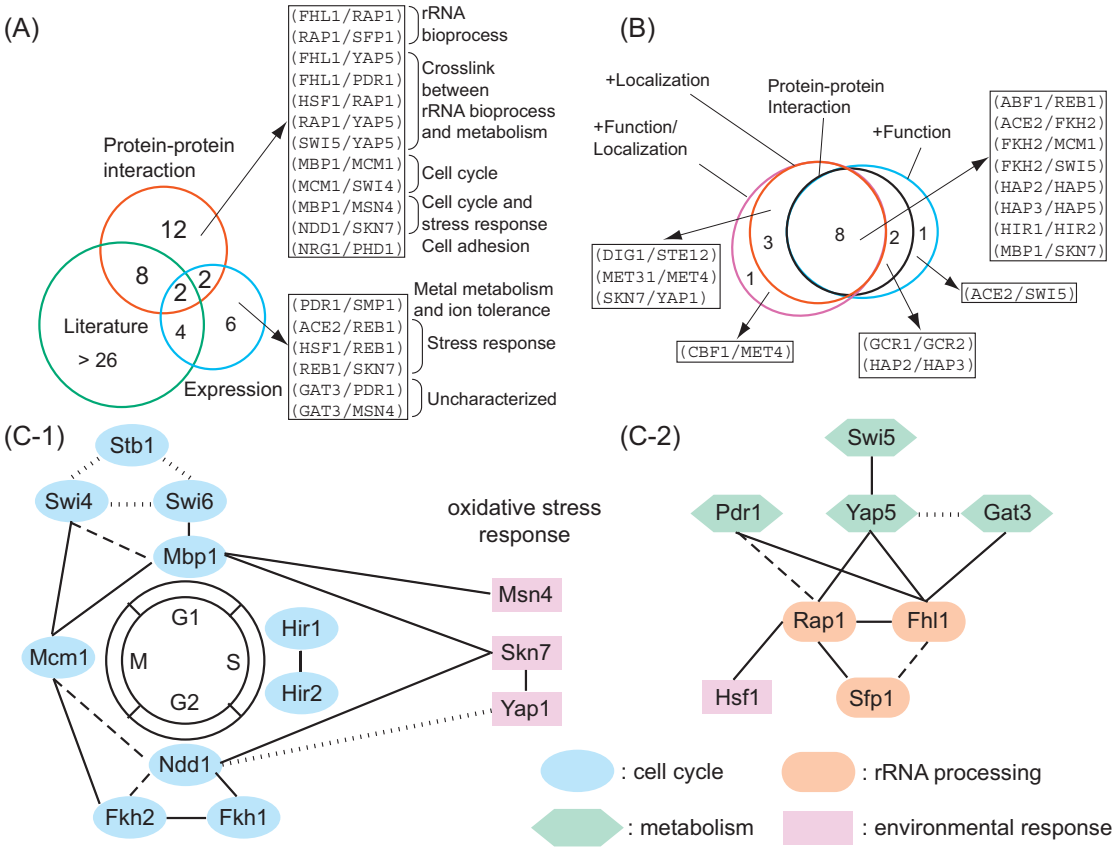


Fig. II.8 (A) Comparison between predictions by protein-protein interaction data (with $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 3$), and those by expression data (Banerjee and Zhang, 2003). All TF pairs listed in this diagram are those less than P-value cutoff (that is, P-value< 0.05) with Holm's correction. Predictions made only by using protein-protein interaction or only by expression, and their possible functions are shown. The number of predictions by both data depends on parameters and expression or protein-protein interaction data to be used, and it may fluctuate. (B) The effect of integrating various biological data. The number of true-positive predictions in each condition, with $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 2$, $P_{mw} \leq 0.05$, and the mutual information threshold $I_{\rm loc} = I_{\rm fnc} = 0.01$, are shown. The total number of predictions is 34 in using only protein-protein interaction data, and 37 in the other conditions. (C) Predicted cooperative TF clusters. (C-1) The cluster of TFs involved in cell cycle. The allocations of TFs to four cell-cycle phases are determined based on Simon *et al.*, 2001. (C-2) The cluster of TFs related to rRNA processing. Marks surrounding a protein show a biological process that the protein contributes to. A broken line indicates cooperativity detected as a triad or a module, and a dotted one does weak cooperativity in a triad or a module.

Table. II.1 Predicted cooperative TF pairs (P_B [a] < 0.001, $O_{\min}$ [b] $= 3$, $I_{\min}$ [c] $= 3$, P_{mw} [d] ≤ 0.05)

	TF1	TF2	P_{mw} (vs TF1)	P_{mw} (vs TF2)	Literature evidence	Expression data[e]
1	HIR1	HIR2	5.41E-07	4.14E-07	Loy *et al.*, 1999	Y[f]
2	FHL1	RAP1	3.48E-04	8.10E-34	NA	N
3	MCM1	SWI4	6.53E-04	2.57E-04	NA	N
4	RAP1	YAP5	1.41E-03	1.43E-10	NA	N
5	FKH1	NDD1	2.09E-03	9.63E-04	Kumar *et al.*, 2000	y[g]
6	FHL1	PDR1	2.29E-03	2.30E-06	NA	N
7	FKH2	MCM1	3.99E-03	3.02E-04	Spector *et al.*, 1997	Y
8	MBP1	MSN4	4.23E-03	2.88E-03	NA	N
9	ARG80	ARG81	5.61E-03	2.92E-04	Mamnun *et al.*, 2002	y
10	NRG1	PHD1	6.35E-03	8.02E-05	NA	N
11	RAP1	SFP1	6.42E-03	1.15E-03	NA	N
12	FHL1	GAT3	6.66E-03	3.51E-08	NA	Y
13	FHL1	YAP5	7.52E-03	6.77E-18	NA	N
14	HAP2	HAP5	2.51E-04	1.01E-02	McNabb *et al.*, 1995	N
15	HAP3	HAP5	7.96E-04	1.01E-02	McNabb *et al.*, 1995	N
16	NRG1	YAP6	4.58E-07	1.17E-02	NA	Y
17	MBP1	MCM1	1.17E-02	2.31E-04	NA	N
18	FKH1	FKH2	3.88E-03	1.43E-02	Spector *et al.*, 1997; Koranda *et al.*, 2000	y
19	HSF1	RAP1	1.44E-02	1.97E-02	NA	N
20	SWI5	YAP5	2.25E-02	1.17E-02	NA	N
21	NDD1	SKN7	3.09E-02	1.42E-02	NA	N
22	MBP1	SKN7	3.59E-02	1.80E-05	Bouquin *et al.*, 1999	N
23	GCR1	GCR2	4.21E-02	2.41E-02	Uemura *et al.*, 1997; Turkel and Bisson, 1999	N
24	SKN7	YAP1	8.15E-03	4.34E-02	Lee *et al.*, 1999	N

[a] P_B, P-value for TF binding to chromatin as described in Lee *et al.*, 2002.
[b] $O_{\min}$, threshold for the number of genes in overlap sets. TF pairs whose overlap set has more than $O_{\min}$ genes are considered.
[c] $I_{\min}$, threshold for the number of interactions. $D(i, j, 2)$ is calculated for proteins that have more than $I_{\min}$ interactions.
[d] P_{mw}, P-value as a result of Mann-Whitney U test.
[e] predictions by using gene expression data in Banerjee and Zhang, 2003 ($P_B < 0.001$).
[f] capital "Y" means that P-value< 0.05 with Holm's correction in Banerjee and Zhang, 2003.
[g] small "y" means that P-value< 0.05, but not significant with Holm's correction in Banerjee and Zhang, 2003.

Table. II.2 Predicted cooperative TF triads ($P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 3$, $P_{mw} \leq 0.05$)

TF1	TF2	TF3	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	P_{mw}(vs. TF3)
FHL1	PDR1	RAP1	7.55E-03	8.19E-05	2.02E-05
HAP2	HAP3	HAP5	3.52E-04	5.01E-03	1.01E-02
FHL1	RAP1	YAP5	1.67E-02	9.08E-06	1.14E-12
MBP1	MCM1	SWI4	2.47E-02	7.10E-04	1.48E-03
FHL1	RAP1	SFP1	3.77E-02	2.95E-03	2.08E-03
[a]FHL1	GAT3	RAP1	4.73E-02	1.94E-04	1.37E-02
NDD1	SKN7	YAP1	7.61E-03	7.98E-03	7.81E-02
FKH1	FKH2	NDD1	2.34E-02	9.02E-02	6.19E-03
NRG1	PHD1	YAP6	1.38E-03	1.42E-04	1.16E-01
FHL1	GAT3	YAP5	5.06E-01	8.61E-03	7.67E-05

[a]: P-value boundary upper which all of three Mann-Whitney U tests satisfy the threshold with Holm's correction, and below which two of them satisfy it.

Table. II.3 Predicted cooperative TF modules ($P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 3$, $P_{mw} \leq 0.05$)

TF1	TF2	TF3	TF4	P_{mw} (vs. TF1)	P_{mw} (vs. TF2)	P_{mw} (vs. TF3)	P_{mw} (vs. TF4)
FHL1	RAP1			<2.2E-308	1.12E-18		
PHO2	PHO4			8.59E-11	2.08E-11		
MBP1	SWI6			8.22E-06	8.89E-04		
FHL1	RAP1	YAP5		3.55E-32	<2.2E-308	<2.2E-308	
FKH1	FKH2	NDD1		1.42E-05	7.67E-104	1.18E-30	
[a]FKH2	MCM1	NDD1		9.16E-08	1.22E-04	2.71E-48	
STB1	SWI4	SWI6		6.07E-08	7.56E-02	3.17E-24	
MBP1	MCM1	SWI4	SWI6	5.79E-02	1.07E-34	5.94E-43	5.21E-34

[a]: P-value boundary upper which all of the Mann-Whitney U tests satisfy the threshold with Holm's correction, and below which all but one satisfy it.

Skn7 are known to function in the G_1/S phase (Bouquin *et al.*, 1999). Hir1 and Hir2, in the S phase, contribute to transcriptional repression (Spector *et al.*, 1997).

On the other hand, Arg80/Arg81, Gcr1/Gcr2 and Skn7/Yap1 are involved in biological processes other than cell cycle. Arg80 and Arg81 regulate the metabolism of arginine (Jamai *et al.*, 2002). Gcr1 and Gcr2 contribute to regulating glycolysis (Uemura *et al.*, 1997; Turkel and Bisson, 1999). Skn7 and Yap1 play a role in the oxidative stress response (Lee *et al.*, 1999).

In addition, Pho2/Pho4 and Stb1/Swi4/Swi6 are detected as a module (Table II.3). Pho2 and Pho4 are known to function in the regulation of phosphate metabolism (Berben *et al.*, 1988). Stb1, Swi4 and Swi6 regulate START in the G_1 phase of the cell cycle (Ho *et al.*, 1999).

3.2.2 Newly discovered TF pairs

As for newly discovered TF pairs without literature evidence, we may classify them into three groups: cooperative pairs involved in the cell cycle, those concerned with Nrg1, and those including Fhl1, Rap1 and Yap5.

■Predictions related to cell cycle

Mbp1/Mcm1, Mbp1/Msn4, Mcm1/Swi4 and Ndd1/Skn7 are thought to be involved in the cell cycle (Fig. II.8C-1). Particularly, we infer that Mcm1/Swi4 and Mbp1/Mcm1 cooperatively function in the M/G_1 phase of the cell cycle. In the M/G_1 phase, it is argued that Mcm1 and Swi4 form a feedforward loop, in which Mcm1 activates Swi4, and both of them activate Clb2 (Lee *et al.*, 2002). Thus, it is very reasonable to conclude that Mcm1 and Swi4 are cooperative. On the other hand, Mbp1 is known to activate Swi4 (Simon *et al.*, 2001) and may participate in the feedforward loop. The cooperativity of Mbp1/Mcm1/Swi4 and Mbp1/Mcm1/Swi4/Swi6, in which Mbp1/Swi6 and Swi4/Swi6 are known to be cooperative (Koch *et al.*, 1993), supports this.

In Mbp1/Msn4 and Ndd1/Skn7, Mbp1 and Ndd1 regulate a progression from the G phase in the cell cycle (G_1 to S and G_2 to M, respectively) (Simon *et al.*, 2001). On the other hand, Msn4 and Skn7 are involved in the response to the oxidative stress (Hasan *et al.*, 2002; Morgan *et al.*, 1997). In addition, in Ndd1/Skn7/Yap1, Skn7/Yap1 contributes to the oxidative stress response (Loy *et al.*, 1999). From these, Ndd1/Skn7 and Mbp1/Msn4 may contribute to the mechanism of the cell cycle arrestation by stress.

■Predictions involving Nrg1

As for cooperative pairs concerned with Nrg1, a possible function of Nrg1/Phd1 and Nrg1/Yap6 is a control of cell adhesion. Nrg1 and Phd1 are respectively thought to be involved in the control of cell adhesion, particularly a regulation of Flo11, which is a key protein of cell adhesion (Braus *et al.*, 2003; Gancedo, 2001; Vyas *et al.*, 2003). Relation of Yap6 to cell adhesion is not yet known. However, given a weak cooperativity of Nrg1/Phd1/Yap6 (2 out of 3 Mann-Whitney U tests satisfy the threshold), it is possible that Nrg1, Phd1 and Yap6 cooperatively control the cell adhesion.

■Predictions involving Fhl1, Rap1 and Yap5

The cooperative TF pairs including Fhl1, Rap1 and Yap5 may be classified into two groups: one including Fhl1, Rap1 and Sfp1, and the other consisting of Gat3, Hsf1, Pdr1, Swi5 and Yap5 (Fig. II.8C-2). Fhl1, Rap1 and Sfp1 are individually known to play a role in rRNA processing and ribosome biosynthesis (Fingerman *et al.*, 2003; Hermann-LeDenmat *et al.*, 1994; Miyoshi *et al.*, 2001). Fhl1/Rap1/Sfp1 cooperativity indicates the relations among them.

On the other hand, Gat3, Pdr1, Swi5 and Yap5 are proteins involved in metabolism. Gat3 is known to be responsible for nitrogen metabolism (Cox *et al.*, 1999). Hsf1 regulates the heat shock response (Wiederrecht *et al.*, 1988). Pdr1 participates in

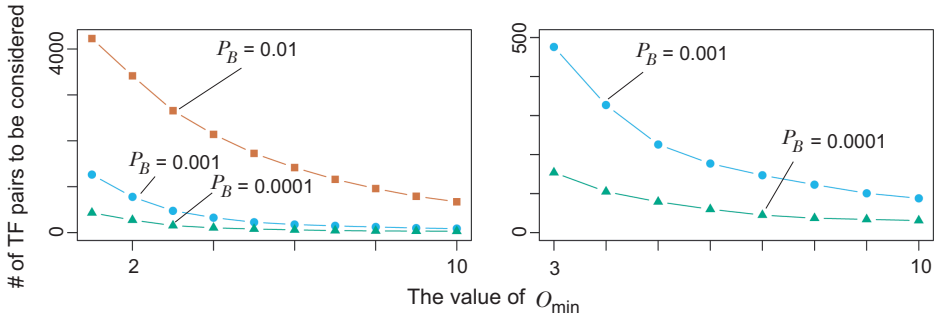


Fig. II.9 Number of TF pairs to be considered on different P_B and $O_{\min}$.

metal metabolism (Mamnun *et al.*, 2002; Tuttle *et al.*, 2003). Swi5 and Yap5 may play a role in some drug metabolism (Butcher and Schreiber, 2004; Mollapour *et al.*, 2004). Hence, according to the relations of TFs controlling rRNA bio-process to those involved in some types of metabolism (including Fhl1/Rap1/Yap5, Fhl1/Gat3/Rap1 and Fhl1/Pdr1/Rap1 cooperativity), we suggest that "some metabolism is regulated rather on translation level than on transcription level".

3.2.3 Effect of parameters

In our study, we use three parameters P_B, $O_{\min}$ and $I_{\min}$. P_B is a threshold to determine proteins regulated by transcription factors (TFs, for short). We consider that, for a TF, proteins whose binding P-value from ChIP data against it is under P_B are regulated by it. $O_{\min}$, with P_B, decides combinations of TFs to be considered (Fig. II.9). For pairs of TFs that have more than $O_{\min}$ proteins regulated by both TFs, we judge their cooperativity. $I_{\min}$ filters proteins in calculating distances based on protein-protein interactions. We calculate distances between two proteins both of which have more than $I_{\min}$ direct interactors.

In Supplementary Tables, "L.E." stands for literature evidence.

■Effect of $O_{\min}$

Predictions results are shown in Supplementary Table A1 through A3.

As we calculate distances between two proteins, $O_{\min}$ must be more than 2. As $O_{\min}$ increases, a number of combinations of TFs to be considered decreases monotonically and so does that of predictions.

In our study, we choose $O_{\min} = 3$ to have as many predictions as possible. It is true that $O_{\min} = 2$ gives the most predictions, but, in case that the overlap set have only two proteins, the central tendency of the set is inevitably represented by one distance and untrustworthy.

■Effect of $I_{\min}$

Predictions results are shown in Supplementary Table A4 through A11.

As the distribution of distances depends on $I_{\min}$, a number of predictions fluctuates with $I_{\min}$. However, a number of predictions that have literature evidence is relatively

stable.

With $P_B = 0.001$ and $O_{\min} = 2$, total $2,266$ proteins are detected with different times to be regulated by some TF to be considered. Among these, 471 proteins have no interactors, 350 have 1, 183 have 2 and 132 have 3. Thus, using $I_{\min} \geq 4$ leads to neglecting more than half of proteins that appear at least once, and is questionable. On the other hand, $I_{\min} = 1$ produces much more predictions than other $I_{\min}$ with the most true predictions and low specificity, and is unfavorable. As, among 350 proteins that have one interactor, interactions of 287 proteins are detected by the yeast two hybrid method, this low specificity may result from the inaccuracy of the yeast two hybrid data.

For these reasons, we use $I_{\min} = 2$ or 3. Particularly, as shown in Fig. 3 in our study, $I_{\min} = 3$ produces the better result.

■Effect of P_B

Prediction results are shown in Supplementary Table A12 through A16.

P_B changes a range of combinations of TFs and their target proteins to be considered. The larger P_B extends the range and makes more predictions. $P_B = 0.01$ produces more than 200 predictions with $O_{\min} = 3$ and $I_{\min} = 3$ (data not shown) and is not handy. On the other hand, predictions made by $P_B = 0.0001$ are reasonable. However, its predictions tend to be confined into those involved in cell cycle. Most predictions contain TFs controlling cell cycle such as ACE2 and SWI4. Therefore, $P_B = 0.001$, we use in our study, is appropriate.

3.2.4 Predictions by using other distance functions

In our study, we use $D(i, j, 2)$, an extended form of Czekanowski-Dice distance ($D_{CD} = D(i, j, 1)$), as a distance function. $D(i, j, l)$ is calculated based on interactors of protein i and j whose D_G from each protein is less than l. D_G is the minimum number of edges needed to traverse from one protein to the other.

Here, we show predictions made by using D_G (Supplementary Table A17 through A19), D_{CD} (Supplementary Table A20 through A24) and $D(i, j, 3)$ (Supplementary Table A25 through A29). For D_{CD} and D_G, we use mean instead of median as the central tendency of a distance set.

As shown in Fig. II.8, $D(i, j, 2)$ gives better results than other distance functions. In addition, $D(i, j, 3)$ is better than D_G and D_{CD}. Therefore, we can conclude that the extended Czekanowski-Dice distance is effective.

3.2.5 Predictions made by using the *in vivo* pull-down data set

In our study, we suggest that predictions made by our method depend on whether targets of TFs can form a complex or not.

However, that is not all that determines results of our method. We show that by making predictions using the *in vivo* pull-down data set, which consists of data on protein complexes.

See Table A30 through A33.

Predictions based on *in vivo* pull-down data are reasonable, but not as good as those

based on data containing interactions from databases and yeast two hybrid method.

Therefore, we deduce that, though formation of complex by targets of TFs is important, other factors, such as interactions between a protein and a complex, also influence predictions.

3.3 Effects of integrating other kinds of biological data

Fig. II.8B shows predictions made by integrating protein localization and function data on conditions of $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 2$, $P_{mw} \leq 0.05$, and $I_{\text{fnc}} = I_{\text{loc}} = 0.01$, the threshold of mutual information that determines which localizations and functions are significantly related to one another. Details of prediction results are provided in Supplementary Table A34 through A36.

As shown in Fig. II.8B, integrating other kinds of biological data enables us to

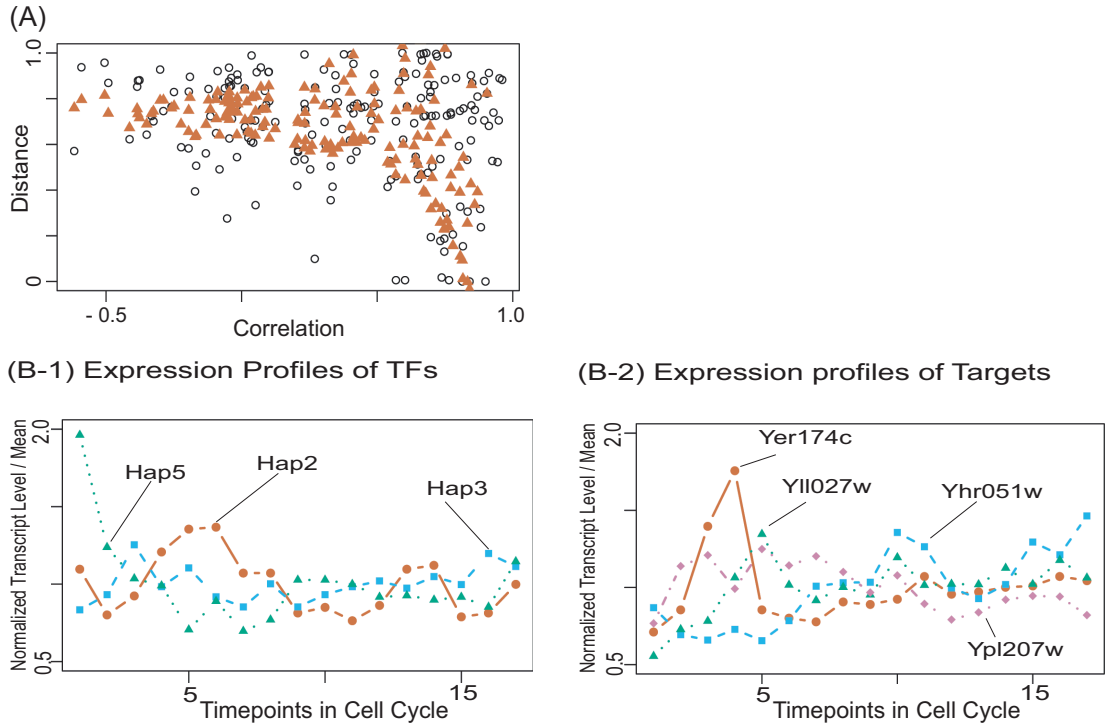


Fig. II.10 (A) Relationships between protein distance $D(i, j)$ and expression correlation coefficient $C(i, j)$ (Pearson correlation coefficient between two expression profiles). For a pair of protein i and j in the overlap sets of TF combinations that are predicted to be cooperative by using protein-protein interaction and expression data, a black circle shows the value of their $C(i, j)$ and $D(i, j)$ on the horizontal and vertical axis respectively. A red triangle denotes approximation of $D(i, j)$ by $C(i, j)$ and number of interactors of protein i and j. (B) Expression profiles of genes concerned with Hap2, Hap3, and Hap5. Expression profiles of three TFs (B-1) and their target genes which all of these three bind to (B-2) in cell cycle according to Cho *et al.*, 1998 are shown. Ylr220w, one of targets of these three TFs, is excluded because it has only one interactor and is thought to be less important.

detect more true-positive cooperative TF pairs. Among these, Dig1/Ste12 regulates the invasive growth (Bardwell *et al.*, 1998), Met31/Met4 is involved in the sulfur amino acid metabolism (Blaiseau *et al.*, 1997), and Cbf1/Met4 in the glutathione metabolism (Wheeler *et al.*, 2003).

While predictions using protein-protein interaction or expression data tend to be related to the cell cycle, integration of various biological data may allow incorporation of many biological aspects and expand the scope of predictions to other biological processes than the cell cycle.

3.4 Relationships between protein-protein interaction-based distance and gene-expression correlation

In the overlap sets of TF pairs that are predicted as cooperative by our method using protein-protein interaction data, we find that the relationship between the interaction-based protein distance and the expression correlation can be approximated by the following equation:

$$D(i,j) = \frac{-aC(i,j) + c(\frac{1}{k(i)} + \frac{1}{k(j)}) + d}{1 - bC(i,j)},$$

where $D(i,j)$ denotes the interaction-based distance between protein i and j, $C(i,j)$ the Pearson correlation coefficient between expression profiles of protein i and j, $k(i)$ the number of interactors of protein i, and a, b, c, d are appropriate constants $(a, b, c, d > 0)$ (Fig. II.10A).

This equation shows that the closeness of two proteins based on our distance measure is proportional to the expression correlation and the profusion of their interactors. As the fact that an expression similarity often implies an existence of protein complexes is already known in (Jansen *et al.*, 2002), we infer that the detection of cooperativity of two TFs by using protein-protein interactions depends on whether the target proteins regulated cooperatively by those TFs can form a complex or not.

4 Discussion

We have proposed a novel method to infer cooperativities of TFs based on protein-protein interactions and shown the biological relevance of our predictions (section 3.2). On the other hand, predictions by our method are a bit limited in coverage. As mentioned in section 3.4, our method may sensitively detect existence of protein complexes. We note that predictions made by our method are not only based on the existence of protein complexes but also capture other biological aspects within protein-protein interactions (see predictions based only on *in vivo* pull-down data, which consist of protein complex data, in supplementary materials). Nevertheless, we found that, though predictions made by using protein-protein interactions and those by using expression data (Banerjee and Zhang, 2003) overlap on already known cooperative TFs, most novel predictions were specific to each method (Fig. II.8A).

This feature shows the possibility that methods using expression data and protein-protein interaction data, each with a bit limited coverage, can complement each other.

Still, there are some advantages of using protein-protein interaction data.

First, it may enable us to detect locally significant correlations among expression profiles. For example, the expression profiles (Cho *et al.*, 1998) of target genes of Hap2/Hap3/Hap5, whose cooperativity is ONLY detected by our method, are not similar as a whole, with no higher correlation coefficient between two expression patterns than 0.3 (Fig. II.10B). However, in timepoints 12-15 of the cell cycle (Cho *et al.*, 1998), where the expression levels of three TFs are similar and they may form a complex to function, the expression profiles of target genes in this short period are rather related to each other, with three of six possible correlation coefficients exceeding 0.7 and five higher than 0.4. The closeness in the protein-protein interaction network may reflect this locally meaningful relatedness of expression patterns.

Second, a method using protein-protein interaction data has some capability for handling post-transcriptional modifications. Though the post-transcript modification is a key factor for many biological phenomena like diseases, the expression data give no information about it. On the other hand, by using interactions between proteins with post-transcript modifications, method using protein-protein interactions can, though indirectly, cooperate post-transcript modifications into its predictions.

Third, the protein-protein interaction network provides a good platform for integrating various biological data. As shown in section 3.3, integration of various data is essential for more comprehensive understanding and prediction of cooperative TFs. Particularly, integration of gene expression data on time course is effective. By selecting time-specific and housekeeping proteins based on expression data and constructing a time-specific protein-protein interaction network from these proteins, our method can be extended to detect time-specific, or dynamic, cooperativity among transcription factors.

Finally, we must note that both predictions made by protein-protein interaction data and by expression-profile data (Banerjee and Zhang, 2003) fairly depend on parameters and datasets, and that the same methods may produce different prediction results by using some elaborately selected dataset. As for datasets, data from high-throughput analyses, like *in vivo* pull-down data which we used, contain a lot of false-positives that should be excluded by exploiting other biological data.

Part III

Statistical protein-chemical interaction prediction

1 Background

In the early stages of the drug discovery process, prediction of the binding of a chemical compound to a specific protein can be of great benefit in the identification of lead compounds (candidates for a new drug). Moreover, the effective screening of potential drug candidates at an early stage generates large cost savings at a later stage of the overall drug discovery process.

In the field of drug discovery, docking analyses and molecular dynamics simulations have been the principal methods used for elucidating the interactions between proteins and small molecules (Shoichet *et al.*, 1992; Jones *et al.*, 1997; Morris *et al.*, 1998; Case *et al.*, 2005). This technique is a 3D-structure based method in which the potential energy for a small molecule to bind to the target protein is evaluated according to a set of equations that model the physical interactions between the receptor and the potential ligand. Because such predictions that are based upon valid free energy calculations are relatively reliable, there are now many docking software tools available such as AutoDock (Morris *et al.*, 1998), DOCK (Shoichet *et al.*, 1992) and GOLD (Jones *et al.*, 1997). However, the requirement of these programs for 3D structural information is a severe disadvantage, as the availability of these data is extremely limited. Although a number of structures in PDB (Berman *et al.*, 2000) is increasing (from 23,642 structures in 2003 to 48,091 structures in 2007), not all proteins which have been derived from many genome-sequencing projects are suitable for experimental structure determination. Hence, the genome-wide application of these methods is in fact not feasible. For example, among the GPCRs (G-protein coupled receptors), whose modulation underlies the actions of 30% of the best-known commercial drugs (Klabunde and Hessler, 2002), the full structure of only a few mammalian members, including bovine rhodopsin (Palczewski *et al.*, 2000) and human beta 2 adrenoreceptor (Rasmussen *et al.*, 2007), is known.

To achieve more comprehensive and faster protein-chemical interaction predictions in the post-genome era producing a vast number of protein sequences whose structural information is not available, it is essential to be able to utilize more readily available biological data and more generally applicable methods which don't require the need for 3D-structural data (Bock and Gough, 2005; Nagamine and Sakakibara, 2007; Jacob and Vert, 2008). In this regard, recent developments in statistical learning and prediction methods hold the promise for very accurate prediction performances when large quantities of learning data are available. In particular, the support vector machine (SVM) statistical method has now been applied to the calculation of putative protein-protein interactions and has been shown to be effective (Bock and Gough,

2001; Gomez *et al.*, 2003; Martin *et al.*, 2005). In addition, the classifications of chemical compounds into drugs and non-drugs using SVM has been proposed (Zernov *et al.*, 2003; Swamidass *et al.*, 2005).

The most prevalent data available for proteins are undoubtedly their amino acid sequences. For chemical compounds, formulas and structures are also generally available in most cases. Moreover, comprehensive metabolite analyses have now been undertaken using mass spectrometry such as CE-MS (Soga *et al.*, 2002), and these have also generated valuable and available data. Based upon these data availabilities, we propose a more comprehensively applicable protein-chemical interaction prediction method than previously described, which is based upon SVM analysis of amino acid sequence data, chemical structure data and mass spectrometry data (Fig. III.1). Unlike the previous approaches to such analyses as described above that assess chemical compounds only and classifying them according to their pharmacological effects, a distinct and novel feature of our proposed approach is the classification of protein and chemical compound pairs into binding and non-binding pairs. We show from our computational experiments that this framework improved the prediction accuracy of the pharmaceutical effects of chemical compounds. Particularly, we demonstrate that our current approach using SVM successfully identified target proteins of chemical compounds that the standard similarity-based methods such as BLAST failed to detect. Another notable feature of our proposed method is the use of mass spectra to encode chemical compounds. In addition, we highlight the effectiveness of using mass spectral data by comparison with and by integrated with existing chemical compound structure data (Fig. III.1).

It is known that interactions of molecules have much more information than the evidence of binding. Protein-protein interactions, for instance, contribute to the elucidation of protein functions (Schwikowski *et al.*, 2000) and transcriptional regulations (Nagamine *et al.*, 2005). Therefore, in comprehensive target protein prediction, we propose the utilization of predicted protein-chemical interactions to describe properties of chemical compounds.

Although the described method showed a relatively high prediction performance (more than 80% accuracy) in cross validation and comprehensive target protein predictions which had approximately twenty thousand prediction targets, it suffered from the problem of predicting many false positives when comprehensive predictions which had millions of prediction targets were conducted. Although these false positives might include some unknown true positives, they were mainly due to the low quality of the negative data, which is one of the common problems in utilizing statistical classification methods such as Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs).

In this study, we describe two strategies, namely two-layer SVM and reasonable negative data design, which are used for the purpose of reducing the number of false positives and improving the applicability of our method for comprehensive prediction. In two-layer SVM, in which outputs produced by the first-layer SVM model are utilized as inputs to the second-layer SVM, in order to design negative data which produce fewer false positives, we iteratively constructed SVM models or

classification boundaries and selected negative sample candidates according to predetermined rules. By using these two strategies, the number of predicted candidates was reduced to around 100 (Table III.9) in experiments in which the potential ligands for some druggable proteins (P10275 (androgen receptor), P11229 (muscarinic acetylcholine receptor M1) and P35367 (histamine H1 receptor)) were predicted on the basis of more than 100,000 compounds in the PubChem Compound database (`http://pubchem.ncbi.nlm.nih.gov/`).

With the aim of validating the usefulness of our method, our proposed prediction model with fewer false positives was applied to the PubChem Compound database in order to predict the potential ligands for the "androgen receptor", which is one of the genes responsible for prostate cancer. We verified some of these predictions by measuring the IC_{50} values in an *in vitro* assay.

Biological experiments, which are conducted to verify the computational predictions based on statistical methods, docking methods or molecular dynamics methods, typically involve success as well as failure. In addition to the fast calculation and wide applicability, one of the merits of using statistical methods involving training with known data is that results obtained by verification experiments can be efficiently utilized or used as feedback to produce new and more reliable predictions. Most previous work on virtual screening has focused on the computational prediction and listing of dozens or hundreds of candidates, followed by their experimental verification. However, only on rare occasions have these experimental results been utilized for the further improvement of computational predictions and experiments. Moreover, even without verification experiments, additional data can be acquired from, for example, relevant literature and used for enhancing the prediction reliability.

Therefore, we propose a strategy for the effective combination of computational prediction and experimental verification. Our second computational prediction utilizing feedback from the first experimental verification which successfully discovered novel ligands (Fig. III.16 and III.18B). Our approach suggests the significance of utilizing statistical learning methods and feedback from experimental results in the drug lead discovery.

2 Methods
In silico experiment section

2.1 Sample representation

For a protein-chemical compound pair, the protein is represented by its amino acid sequence and the compound is denoted by either its mass spectrum or its chemical structure. The combination of a feature vector for a protein and that for a chemical compound constitutes a sample.

Table III.1 Positive and negative samples of ADR drugs dataset containing metazocine

drug	target	attribution
metazocine	α_{1A}	positive
metazocine	α_{1B}	positive
metazocine	α_{1D}	positive
metazocine	α_{2A}	negative
metazocine	α_{2B}	negative
metazocine	α_{2C}	negative
metazocine	β_1	negative
metazocine	β_2	negative
metazocine	β_3	negative

2.2 Experimental datasets

We constructed two experimental datasets, an adrenergic receptor (ADR) drug and DrugBank dataset.

The ADR drug dataset was based on ARDB (`http://ardb.bjmu.edu.cn/default.htm`) as of February, 2006. and comprises of 48 ADR drugs, including 22 agonists and 26 antagonists, and 9 human adrenergic receptors. Out of the total possible number ($9\times$ 48 = 432) of protein-chemical compound pairs, 142 were found to be positive samples, or interacting protein-chemical pairs (see Supplementary Table A37), and the remaining 290 are considered negative or non-binding protein-chemical pairs. We regarded ADR α_1 targeted drugs as binding to 3 receptors in the ADR α_1 family ($\alpha_{1\mathrm{A}}$, $\alpha_{1\mathrm{B}}$ and $\alpha_{1\mathrm{D}}$). For example, if a drug x is known to bind only to ADR β_1, a pair (x, ADRβ_1) is regarded as positive, and other eight pairs such as (x, ADRβ_2) and (x, ADRα_2A) are treated as negative samples.

For example, metazocine which binds to ADR α_1 constitutes positive and negative samples in Table III.1.

The DrugBank dataset was first constructed from Approved Drug Target Protein Sequences data, downloaded in February, 2006, from the DrugBank database (Wishart *et al.*, 2008). These data consist of 519 approved drugs, whose mass spectra were obtainable from the NIST/EPA/NIH mass spectral library (NIST 05) (`http://www.nist.gov/`) incorporating 190,825 EI (Electron Impact) spectra for 163,198 chemical compounds, and their 291 associated target proteins, constituting 980 interacting pairs (see Supplementary Table A38). In order to examine use of mass spectrometry data, the first DrugBank dataset excluded drugs whose mass spectrometry data were not available. An example within this dataset is the dopamine receptor, COX2, and the sodium-dependent serotonin transporter. In this dataset, n random pairs of drugs and proteins, except for positive pairs, are regarded as negative samples.

The DrugBank dataset was reconstructed from Approved DrugCards data, which was downloaded in February, 2007, from the DrugBank database (Wishart *et al.*, 2008). These data consist of 964 approved drugs and their 456 associated target proteins, constituting 1,731 interacting pairs or positives (see Supplementary Table A39).

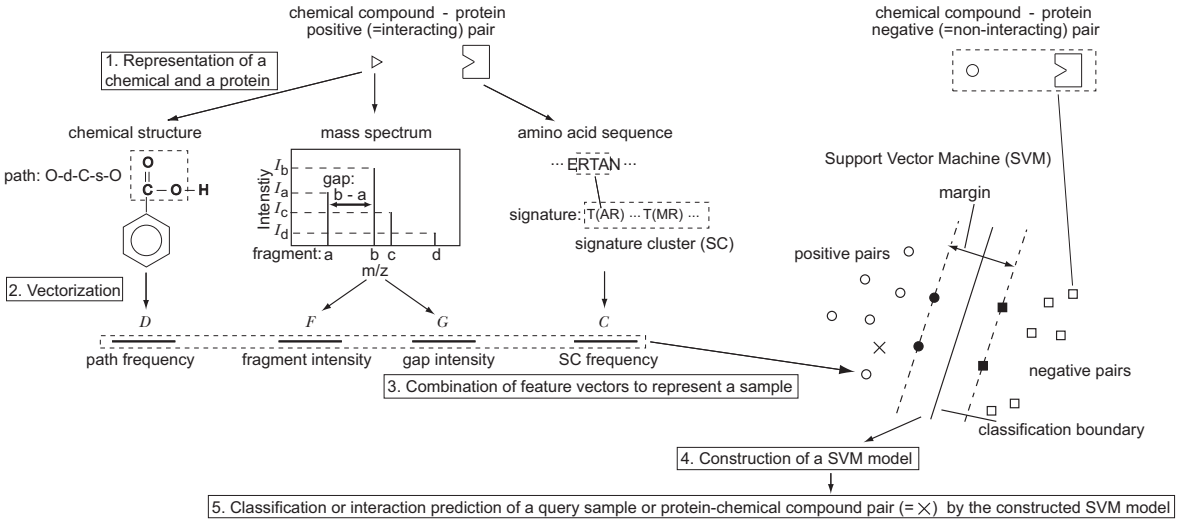


Fig. III.1 Protein-chemical interaction prediction strategy. Schematic illustration of the one-layer SVM system. Both interacting and non-interacting pairs of chemical compounds and proteins are regarded as samples. In step 1, a chemical compound is represented by its mass spectrum or its chemical structure and a protein is characterized by its amino acid sequence. In step 2, the non-numerical data in step 1 are mapped to a numerical feature vector space. See Methods for details. In step 3, feature vector types are selected to represent a sample and in step 4 an SVM model is constructed from the positive and negative pairs. In step 5, a prediction of whether a query sample displayed an interaction or not was made using the SVM model constructed in step 4.

As the second DrugBank dataset (DrugBank2 dataset) was developed to construct prediction models applied to the large public compound databases mainly providing chemical structure data, the DrugBank2 dataset includes drugs without mass spectrometry data.

2.3 Statistical prediction model

Our basic strategy of statistically predicting interactions between proteins and chemical compounds are shown in Fig. III.1. Two main characteristics of our strategy are (i) utilization of the statistical learning method support vector machines (Vapnik, 1998; Cristianini and Sawe-Taylor, 2000) and (ii) use of readily available data including amino acid sequence data, chemical structure data and mass spectrometry data, both of which contribute to realization of comprehensive protein-chemical interaction prediction. In the following sections, these characteristics are explained.

Moreover, we propose the two-layer SVM model consisting of the first-layer SVM models which are constructed following to the basic strategy (Fig. III.1 and the second-layer SVM which utilizes outputs of the first-layer SVM models (Fig. III.2) as inputs.

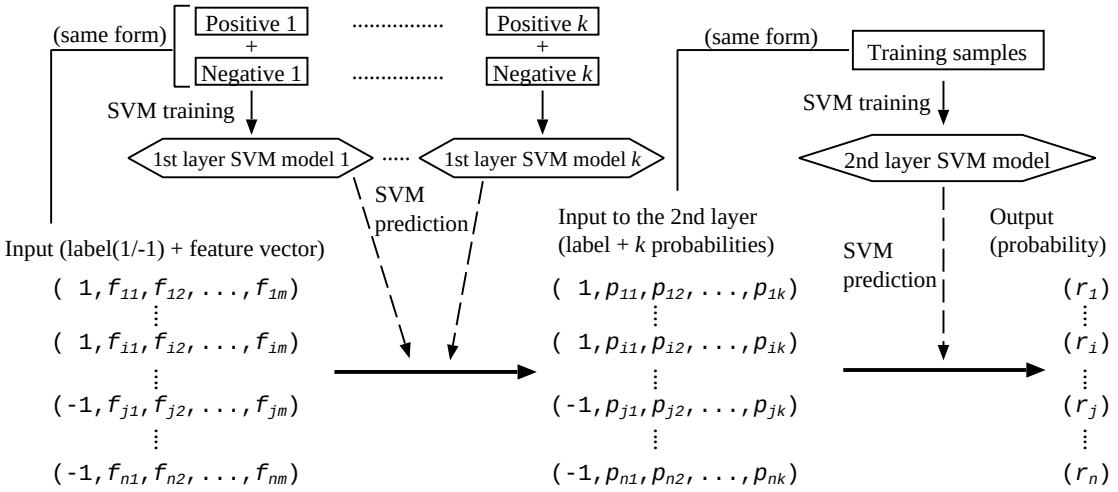


Fig. III.2 Schematic illustration of the two-layer SVM system. It functions as follows.

1. Given $\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}_1 = (f_{11}, f_{12}, \ldots, f_{1l}), \boldsymbol{x}_2 = (f_{21}, f_{22}, \ldots, f_{2l}), \ldots, \boldsymbol{x}_n = (f_{n1}, f_{n2}, \ldots, f_{nl})\}$, where f_{ij} is the value for the jth dimension of the feature vector for the sample, or the protein-chemical pair, i expressed in Eqs (III.5) and (III.8).
2. k first-layer SVM models are applied to $\boldsymbol{\mathcal{X}}$ to produce $\boldsymbol{\mathcal{P}} = \{\boldsymbol{P}_1 = (p_{11}, p_{12}, \ldots, p_{1k}), \boldsymbol{P}_2 = (p_{21}, p_{22}, \ldots, p_{2k}), \ldots, \boldsymbol{P}_n = (p_{n1}, p_{n2}, \ldots, p_{nk})\}$, where p_{ij} is the output of the first-layer SVM model j applied to the sample i and shows the possibility calculated in Eq. (III.2) that i is positive.
3. The second-layer SVM model is applied to $\boldsymbol{\mathcal{P}}$ to produce the final output $\boldsymbol{\mathcal{R}} = (r_1, r_2, \ldots, r_n)$, where r_i is the possibility that the sample i is positive.

2.3.1 Support vector machines

Given n samples, each of which has a m-dimensional feature vector ($\boldsymbol{x}_i = (x_i^1, \ldots, x_i^m)$) and one of two classes such as binding and non-binding ($y_i \in \{1, -1\}$), an SVM produces the classifier

$$f(\boldsymbol{x}) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b^*\right), \tag{III.1}$$

where $\boldsymbol{x}$ is any new object which needs to be classified, $K(\cdot, \cdot)$ is a kernel function which indicates the similarity between two vectors and $(\alpha_1, \ldots, \alpha_n)$ are the learned parameters (Vapnik, 1998; Cristianini and Sawe-Taylor, 2000)

The output of an SVM can be regarded as a probability using the following formula (Platt, 2000),

$$p(y = 1|\boldsymbol{x}) = \frac{1}{1 + \exp\left(A(\sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b^*) + B\right)} \tag{III.2}$$

A and B are parameters given by solving the likelihood maximization.

Supplementary explanation for SVM is provided in Appendix B.

In this report, the *LIBSVM* 2.81 (Chang and Lin, 2001) program was employed to construct the SVM model.

2.3.2 Feature representation

In order to apply statistical methods, including the support vector machine expressed in Eq. (III.1), to non-numerical data such as character strings, this type of data must be converted into some numerical data. The feature representation is one way to realize this in which it is evaluated whether a feature, such as a specific character in strings, exists in a sample or how many times a feature appears in a sample. As a result of the feature representation, a non-numerical sample is converted into a numerical vector, or a feature vector, whose ith value corresponds to the existence, the frequency or the value of the ith feature considered. Many statistical methods, including SVM, utilize the similarity between feature vectors to solve the problem.

■Protein description

We define "description" as mapping the non-numerical data like amino acid sequences into an n-dimensional numerical vector space so that these data can be utilized in the statistical learning.

There are two approaches to encode amino acid sequences that have been mainly used in the previous works.

1. Use of the physico-chemical properties of amino acids (Bock and Gough, 2001; Yanover and Hertz, 2005).
2. Use of amino acid subsequences (Bhasin and Raghava, 2004; Martin *et al.*, 2005; Xiao *et al.*, 2006; Yu *et al.*, 2006).

Bock and Gough utilized (1) charge, (2) hydrophobicity, (3) surface tension of each amino acid in the sequence to convert strings into numeric data. Given a protein A with L residues, M features (ex. hydrophobicity) considered and some fixed length K, their general encoding formula is as follows,

$$\{\boldsymbol{y}\}^i = f(\{\boldsymbol{v}\}^i),\ i \in 1, \ldots, M,\ \boldsymbol{y} = (y_1, \ldots, y_K),\ \boldsymbol{v} = (v_1, \ldots, v_L)$$

$$\{\phi_A\} = \{\boldsymbol{y}\}^1 \oplus \{\boldsymbol{y}\}^2 \oplus \cdots \oplus \{\boldsymbol{y}\}^M.$$

Here, $\{\boldsymbol{v}_j\}^i$ is a value for the feature i of the j th amino acid in the protein A. A function $f : \mathcal{R}^L \to \mathcal{R}^K$ transforms a L dimensional feature vector $\{\boldsymbol{v}\}$ into a normalized K dimensional vector $\{\boldsymbol{y}\}$ to arrange widely varying sequence length. $\oplus$ means concatenation of vectors.

For example, a 4-letter amino acid sequence $ACDE$ can be represented with 3 features charge (c), hydrophobicity (h) and surface tension (s) and without the transformation function f as follows.

$\phi(ACDE) = (c(A),\ c(C),\ c(D),\ c(E),\ h(A),\ h(C),\ h(D),\ h(E),\ s(A),\ s(C),\ s(D),\ s(E))$.

Here, $c(A)$, $h(A)$ and $s(A)$ represent charge, hydrophobicity and surface tension for alanine, or A, respectively.

On the other hand, Bhasin *et al.* used dipeptide composition to describe a protein sequence. Their encoding approach is as follows,

$$\text{Fraction of } dep(i) = \frac{\text{Total number of } dep(i)}{\text{Total number of all possible dipeptides}}. \tag{III.3}$$

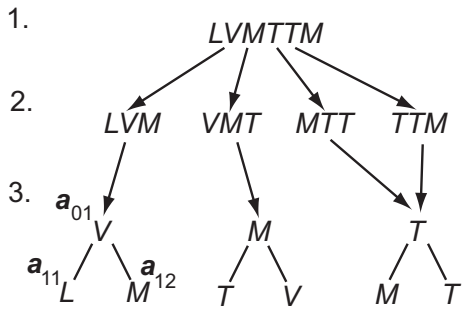


Fig. III.3 Signature (Martin *et al.*, 2005)

As we have 20 amino acids, an amino acid sequence is converted into a $20 \times 20 = 400$ dimensional feature vector.

For example, a 4-letter amino acid sequence $ACDE$ can be represented as follows,

$$\begin{array}{rc} & \begin{array}{ccccccccc} 1:AA & 2:AC & 3:AD & & CD & & DE & & YY \end{array} \\ \phi(ACDE) = (& \begin{array}{ccccccccc} 0, & 1/3, & 0, & \ldots, & 1/3, & \ldots, & 1/3, & \ldots, & 0 \end{array}) \end{array}$$

Both of these approaches worked as the binding including protein-protein interaction or protein-peptide interaction is a product of physico-chemical activities of atoms and a relatively local event that can be represented by small subsequences. In addition, amino acid subsequences can represent localization or structure of proteins (Xiao *et al.*, 2006; Yu *et al.*, 2006).

However, there are some defects of these approaches. For example, both of the two approaches easily have a large number of features. If tetrapeptide composition is considered, a feature vector has 20^4 =160,000 features. When 100 features are considered for each amino acid, a protein with 300 residues can be represented as a 30,000 dimensional feature vector. As the computation time of SVM is $O(nm)$, or proportional to the product of a number of samples n and a number of features m, a lower dimensional feature vector is preferable.

Based on these merits and defects, we, in our current study, proposed a method for protein description with the following features.

1. Integration of physico-chemical property and amino acid subsequence approaches.
2. Reduction of features.

First, we consider 3-mer signature used in (Martin *et al.*, 2005), which is one of amino acid subsequence methods. Signature is generated as follows (Fig. III.3).

1. Consider a full length amino acid sequence.
2. Extract all possible continuous 3 amino acids from an amino acid sequence.
3. Regard an amino acid 3-mer as a binary tree with an amino acid in the middle of the 3-mer (a_{01}) as a root and the others (a_{11} and a_{12}) as a leaf. Here, 3-mer $a_{11}a_{01}a_{12}$ and $a_{12}a_{01}a_{11}$ produce a same signature like MTT and TTM in Fig III.3.

Table III.2 Amino acids descriptors by Venkatarajan and Braun, 2001

amino acids	Eigenvector : E_i (Eigenvalue : λ_i)				
	E_1(1961.504)	E_2(788.2)	E_3(539.776)	E_4(276.624)	E_5(244.106)
A	0.008	0.134	-0.475	-0.039	0.181
R	0.171	-0.361	0.107	-0.258	-0.364
N	0.255	0.038	0.117	0.118	-0.055
D	0.303	-0.057	-0.014	0.225	0.156
C	-0.132	0.174	0.07	0.565	-0.374
Q	0.149	-0.184	-0.03	0.035	-0.112
E	0.221	-0.28	-0.315	0.157	0.303
G	0.218	0.562	-0.024	0.018	0.106
H	0.023	-0.177	0.041	0.28	-0.021
I	-0.353	0.071	-0.088	-0.195	-0.107
L	-0.267	0.018	-0.265	-0.274	0.206
K	0.243	-0.339	-0.044	-0.325	-0.027
M	-0.239	-0.141	-0.155	0.321	0.077
F	-0.329	-0.023	0.072	-0.002	0.208
P	0.173	0.286	0.407	-0.215	0.384
S	0.199	0.238	-0.015	-0.068	-0.196
T	0.068	0.147	-0.015	-0.132	-0.274
W	-0.296	-0.186	0.389	0.083	0.297
Y	-0.141	-0.057	0.425	-0.096	-0.091
V	-0.274	0.136	-0.187	-0.196	-0.299

For example, when signature is directly used, a four-letter acid peptide sequence $ACDE$ can be represented as follows,

$$\begin{array}{rccccccccc} & 1:A(AA) & 2:A(AC) & & C(AD) & & D(CE) & & 4200:Y(YY) \\ \phi(ACDE) = (& 0, & 0, & \ldots, & 1/2, & \ldots, & 1/2, & \ldots, & 0 &) \end{array}$$

For reduction of a number of features, one possible way is to apply clustering to 4,200 signatures. Here, in order to apply statistical clustering methods to non-numerical data, a non-numeric sample needs to be converted into a numeric vector. To realize this, we used physico-chemical properties of each amino acid and integrated the two approaches as mentioned above.

In this study, each amino acid was represented by using the 5 dimensional vector derived from (Venkatarajan and Braun, 2001), which was also used in (Yanover and Hertz, 2005). The feature vector for an amino acid was calculated from eigen values and eigen vectors (Table III.2), which was derived from the principal component analysis applied to 237 kinds of physico-chemical propertied for 20 amino acids. According to (Venkatarajan and Braun, 2001), an amino acid i is represented as follows.

$$\boldsymbol{\alpha}(i) = \left(\sqrt{\lambda}_{\mu=1} E_i^{\mu=1}, \sqrt{\lambda}_{\mu=2} E_i^{\mu=2}, \ldots, \sqrt{\lambda}_{\mu=5} E_i^{\mu=5}\right)$$

For example, alanine (A) can be represented as follows,

$$\boldsymbol{\alpha}(A) = (\sqrt{1961.5} \cdot 0.008,\ \sqrt{788.2} \cdot 0.134,\ \sqrt{539.8} \cdot -0.475,\ \sqrt{276.6} \cdot -0.039,\ \sqrt{244.1} \cdot 0.181)$$

Then, a signature was represented as follows using feature vectors for amino acids,

$$\boldsymbol{\alpha}_s(a_{01}, a_{11}, a_{12}) = \boldsymbol{\alpha}(a_{01}) + \frac{1}{2}\left(\frac{\boldsymbol{\alpha}(a_{11}) + \boldsymbol{\alpha}(a_{12})}{2}\right). \tag{III.4}$$

It can be generally expressed as follows,

$$\boldsymbol{\alpha}_s = \sum_{h=0}^{H} \frac{1}{h+1} \left(\frac{1}{2^h} \sum_{k=1}^{2^h} \boldsymbol{\alpha}(a_{hk}) \right).$$

We applied the variational Bayesian mixture modelling implemented in program R package vabayelMix (Teschendorff *et al.*, 2005) to 4,200 signatures that were represented in the Eq. (III.4). Although the variational Bayesian mixture modelling was used in this study, other clustering methods such as hierarchical clustering can be applied.

As a result of clustering, 4,200 signatures were clustered into 199 clusters (Supplementary Table A40). In detail, 4,200 signatures was clustered into 34 groups at first. Then, the same clustering method was applied to signatures in each group to produce 199 clusters.

In our current study, according to these 199 clusters (see Supplementary Table A40), a feature vector for protein p, $C(p)$, is calculated as follows,

$$C(p) = (\rho_p(c))_{c \in \mathcal{C}}, \quad \rho_p(c) = \begin{cases} \dfrac{f_p(c)}{\sum_{i \in \mathcal{C}(p)} f_p(i)} & \text{if } c \in \mathcal{C}(p) \\ 0 & \text{otherwise} \end{cases} \tag{III.5}$$

For example, in this study, a four-letter acid peptide sequence $ACDE$ can be represented as follows,

$$C(ACDE) = \begin{pmatrix} {}^{1}c(C(AD)) & 2 & {}^{148}c(D(CE)) & 199 \\ 1/2, & 0, \ldots, & 1/2, & \ldots, \ 0 \end{pmatrix}$$

Here, the numbering such as 1 and 148 indicates the numbering of clustering in Supplementary Table A40. A signature $C(AD)$ belongs to the 1st cluster in Supplementary Table A40.

■Chemical compound description by mass spectrometry data

In this study, two types of feature vectors, fragment vector $F(c)$ and gap vector $G(c)$, are produced from the mass spectrometry data showing m/z values and intensities for each m/z value, which are scaled 1 to 999 in a chemical compound.

A fragment vector for a chemical c, $F(c)$, is defined as follows,

$$F(c) = (\phi_c(m))_{m \in \mathcal{M}}, \quad \phi_c(m) = \begin{cases} I_c(m) & \text{if } m \in \mathcal{M}(c) \\ 0 & \text{otherwise} \end{cases} \tag{III.6}$$

where $\mathcal{M}$ is a set of m/z values that appear at least once in mass spectra in the dataset and $\mathcal{M}(c)$ are those found at least once in a chemical c. $I_c(m)$ is the intensity of an m/z value m in c (Fig. III.4).

For example, when we have mass spectra for Compound 1 (C_1) and Compound 2 (C_2) in Fig. III.4, they can be represented as follows,

$$F(C_1) = \begin{pmatrix} (m/z = 40) & (120) & (160) & (200) \\ 50, & 10, & 0, & 100 \end{pmatrix}$$

$$F(C_2) = \begin{pmatrix} 80, & 0, & 140, & 0 \end{pmatrix}$$

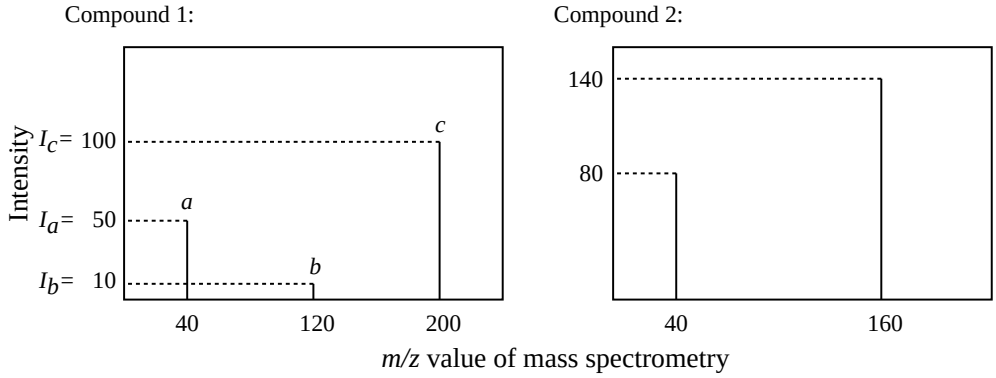


Fig. III.4 Mass spectrometry data

A gap is defined between two peaks, or m/z values, and reflects the substructure that is represented by the bigger m/z value and not by the smaller m/z value. A gap vector, $G(c)$, is calculated by

$$G_t^w(c) = (\xi_c(m))_{m\in\mathcal{M}_g, m\geq w}, \quad \xi_c(m) = \begin{cases} gap_c(m) & \text{if } m \in \mathcal{M}_g(c) \\ 0 & \text{otherwise} \end{cases}$$
$$gap_c(m) = \sum_{i;i+m\in\mathcal{M}(c), I_i\geq t} g_i(m), \quad g_i(j-i) = I_i \times \frac{\ln(I_j)}{\sum_{k;k>i, I_k\geq t} \ln(I_k)} \tag{III.7}$$

where $\mathcal{M}_g$ is a set of gaps greater than the threshold w that appear at least once in the mass spectra of compounds in the dataset, and $\mathcal{M}_g(c)$ is the set of gaps found at least once in a chemical c (Fig. III.4). w is calculated to exclude small gaps for which there are no possible structures. $g_i(j-i)$ is a virtual intensity for a gap that can be produced by the breakdown of a structure represented by m/z value j to that of i. Herein, I_i is the intensity for m/z value i and t is a denoising threshold for these intensities. $\mathcal{M}(c)$ is as defined in Eq. (III.6).

For example, for Compound 1 (C_1) in Fig. III.4, $g_a(c-a)$ is calculated as follows,

$$g_a(c-a) = g_{40}(160) = \begin{cases} I_a \times \dfrac{\ln(I_c)}{\ln(I_b)+\ln(I_c)} = 50 \times \dfrac{\ln(100)}{\ln(10)+\ln(100)} = 33.3 & \text{if } t < 10 \\ I_a \times \dfrac{\ln(I_c)}{\ln(I_c)} = 50 \times \dfrac{\ln(100)}{\ln(100)} = 50 & \text{otherwise.} \end{cases}$$

When $t \leq 10$, $gap_{C_1}(80)$ is calculated as follows,

$$\begin{aligned} gap_{C_1}(80) &= g_a(b-a) + g_b(c-b) = g_{40}(80) + g_{120}(80) \\ &= 50 \times \frac{\ln(10)}{\ln(10)+\ln(100)} + 10 \times \frac{\ln(100)}{\ln(100)} = 26.7. \end{aligned}$$

When mass spectra for Compound 1 (C_1) and Compound 2 (C_2) in Fig. III.4 and $t \leq 10$ are given, they can be represented as follows,

$$\begin{aligned} G_0^{12}(C_1) &= \overset{80}{(}gap_{C_1}(80), \quad \overset{120}{0}, \quad gap_{C_1}(160)\overset{160}{)} = (26.7 \quad 0, \quad 33.3) \\ G_0^{12}(C_2) &= (\quad 0, \quad gap_{C_2}(120), \quad 0 \quad) = (\ 0, \quad 40, \quad 0\ ) \end{aligned}$$

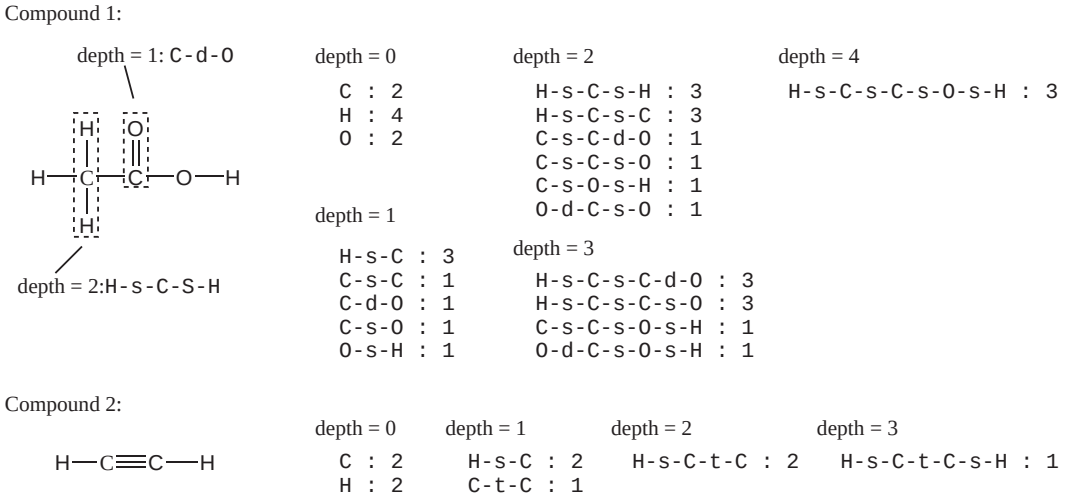


Fig. III.5 Paths extracted from chemical structures.

A virtual intensity g in Eq. (III.7) is based on the following idea. In Fig III.4, a fragment ion a can be derived from fragment ions b and c. In other words, bigger ions b and c break down into a smaller ion a while producing desorption ions whose m/z is $b-a$ or $c-a$. When b and c are regarded as "start", and a as "goal", it is assumed that the possibility that a becomes the "goal" is relevant to I_a, the amount of existence of a, and that the possibility that b and c become the *start* is also relevant to I_b and I_c respectively.

■Chemical compound description by chemical structures

Substructures, or paths, extracted from chemical structures, can be an effective descriptor for chemical compounds when the 2D structures of chemicals are regarded as a graph with an atom as a node and a bond as an edge (Merlot *et al.*, 2003; Clark, 2005; Swamidass *et al.*, 2005).

In this study, we followed the method described in (Swamidass *et al.*, 2005), and a feature vector based on the 2D structure is thus defined as follows,

$$D_l^h(c) = (\psi_c(p))_{p\in\mathcal{P}_l^h}, \quad \psi_c(p) = \begin{cases} \dfrac{f_c(p)}{\sum_{i\in\mathcal{P}_0^h(c)} f_c(i)} & \text{if } \boldsymbol{p}\in\mathcal{P}_l^h(c) \\ 0 & \text{otherwise} \end{cases} \tag{III.8}$$

where $\mathcal{P}_l^h$ is a set of paths whose depth, or a number of bonds within, is between l and h ($h \geq l$) and which appears at least once in chemical structures in the dataset and $\mathcal{P}_l^h(c)$ is that found at least once in a chemical c. $f_c(p)$ is a number of appearances of path p in the structure of chemical compound c.

In Fig. III.5, for Compound 1 (CH_3COOH) and Compound 2 (C_2H_2), D_0^1 is calcu-

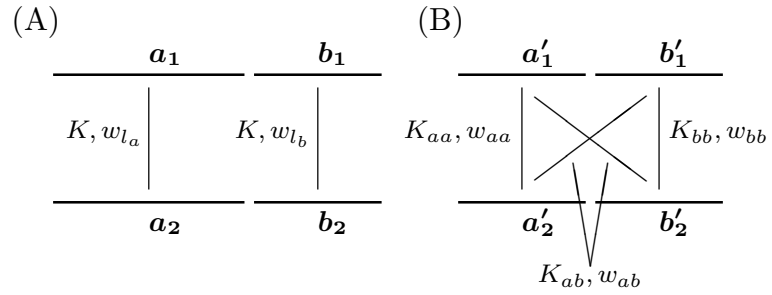


Fig. III.6 The strategy of describing bindings: (A) concatenation. (B) combination

lated as follows.

$$D_0^1(\mathrm{CH_3COOH}) = \overset{\text{C}}{(2/36,}\ \overset{\text{H}}{4/36,}\ \overset{\text{O}}{2/36,}\ \overset{\text{H-s-C}}{3/36,}\ \overset{\text{C-s-C}}{1/36,}\ \overset{\text{C-t-C}}{0,}\ \overset{\text{C-s-O}}{1/36,}\ \overset{\text{C-d-O}}{1/36,}\ \overset{\text{O-s-H}}{1/36}\ )$$

$$D_0^1(\mathrm{C_2H_2}) = (2/10,\ 2/10,\ 0,\ 2/10,\ 0,\ 1/10,\ 0,\ 0,\ 0\ )$$

D_2^3 is represented in the same way as follows.

$$D_2^3(\mathrm{CH_3COOH}) = \overset{\text{HsCsH}}{(\frac{3}{36},}\ \overset{\text{HsCsC}}{\frac{3}{36},}\ \overset{\text{HsCtC}}{0,}\ \overset{\text{CsCdO}}{\frac{1}{36},}\ \overset{\text{CsCsO}}{\frac{1}{36},}\ \overset{\text{CsOsH}}{\frac{1}{36},}\ \overset{\text{OdCsO}}{\frac{1}{36},}\ \overset{\text{HsCtCsH}}{0,}\ \overset{\text{HsCsCdO}}{\frac{3}{36},}\ \overset{\text{HsCsCsO}}{\frac{3}{36},}\ \overset{\text{CsCsOsH}}{\frac{1}{36},}\ \overset{\text{OdCsCsH}}{\frac{1}{36}}\ )$$

$$D_2^3(\mathrm{C_2H_2}) = (0,\ 0,\ \frac{2}{10},\ 0,\ 0,\ 0,\ 0,\ \frac{1}{10},\ 0,\ 0,\ 0,\ 0\ )$$

■Representation of a protein-chemical interaction

In our current study, a sample, or a pair of a protein and chemical compound, is represented by several types of feature vectors; C, D, F and G expressed in Eq. (III.5), (III.6), (III.7), and (III.8).

A straightforward way to represent protein-chemical bindings is to concatenate feature vectors for proteins and compounds, and then to treat a produced vector as one feature vector. For example, a sample S_1, an interaction between peptide $ACDE$ and $\mathrm{CH_3COOH}$ and a sample S_2, an interaction between $ACDE$ and $\mathrm{C_2H_2}$, can be represented as follows,

$$S_1 = (C(ACDE), D_0^1(\mathrm{CH_3COOH})) = \overset{{}^1c(C(AD))}{(1/2,}\ \ldots,\ \overset{{}^{148}c(D(CE))}{1/2,}\ \ldots,\ \overset{199}{0,}\ \overset{\text{C}}{2/36,}\ \ldots,\ \overset{\text{C-t-C}}{0,}\ \ldots,\ \overset{\text{O-s-H}}{1/36}\ )$$

$$S_2 = (C(ACDE), D_0^1(\mathrm{C_2H_2})) = (1/2,\ \ldots,\ 1/2,\ \ldots,\ 0,\ 2/10,\ \ldots,\ 1/10,\ \ldots,\ 0\ )$$

When the RBF kernel $K(S_1, S_2) = \exp(-\gamma\|S_1 - S_2\|^2)$, one of the most frequently used kernel functions, is utilized in Eq. (III.1), the concatenation, $S_1 = (a_1, b_1)$ or $S_2 = (a_2, b_2)$, means that the similarity between a_1 and a_2 and that between b_1 and b_2 are independently evaluated by the same measure and then multiplied to give the overall similarity due to $K(S_1, S_2) = K(a_1, a_2) \cdot K(b_1, b_2)$ (Fig. III.6A). In addition, as the same parameter γ is used in concatenation, $K(a_1, a_2)$ and $K(b_1, b_2)$ have potential weights proportional to the length of feature vector type a or b.

However, it may well be the case that the appropriate measure to evaluate the similarity for one feature vector type differs from that for another feature vector type.

Moreover, to represent and predict protein-chemical interactions, combination effects of different feature vector types can be significant. Different feature vector types and combination effects can also have different weights in integration to evaluate overall similarity of samples.

Therefore, the following formula to determine similarities between two samples in Eq. (III.1) was used,

$$\begin{aligned} K_{\text{int}}(S_1, S_2) &\equiv \textstyle\prod_{I,J \in \boldsymbol{V}} K_{IJ(=JI)}(I_1, J_2) \\ K_{IJ(=JI)}(\boldsymbol{x}, \boldsymbol{y}) &= \begin{cases} \gamma_{IJ}(\boldsymbol{x}^t \boldsymbol{y} + 1)^3 \\ \exp(-\gamma_{IJ} \|\boldsymbol{x} - \boldsymbol{y}\|^2) \\ \tanh(\gamma_{IJ} \boldsymbol{x}^t \boldsymbol{y} + 1) \\ 1 \end{cases} \end{aligned} \tag{III.9}$$

where $\boldsymbol{V}$, for example (a, b), is a set of feature vector types chosen to constitute samples S_1 and S_2 (for example, $S_1 = (a_1, b_1)$) (Fig. III.6B). K_{IJ}, one of four functions in Eq. (III.9), and a parameter γ_{IJ} for a pair of feature vector types are empirically selected to give maximum accuracy. By using different kernels and parameters, each feature vector type is evaluated by different measures, and potentially has appropriate weights that are not proportional to length of vector types. In order to obtain proper inner products, the dimensions, or the number of features in different feature vector types, need to be equivalent.

For example, when feature vector types a and b are considered (Fig. III.6), Eq. (III.9) is as follows,

$$K_{\text{int}}(S_{a'_1 b'_1}, S_{a'_2 b'_2}) \equiv K_{aa}(\boldsymbol{a}'_1, \boldsymbol{a}'_2) \cdot K_{bb}(\boldsymbol{b}'_1, \boldsymbol{b}'_2) \cdot K_{ab}(\boldsymbol{a}'_1, \boldsymbol{b}'_2) \cdot K_{ab}(\boldsymbol{b}'_1, \boldsymbol{a}'_2),$$

where two vector types a' and b' have the same length, or the same number of dimensions ($|a'| \leq |a|,\ |b'| \leq |b|$).

In our current study, to calculate proper inner product, the features are ordered according to the mean squared error calculated among all of the different proteins or chemical compounds in the dataset.

In concrete, for C, D and F in Eq. (III.5), (III.8) and (III.6), features are ordered in the descending order according to $\text{MSE}_i^{F(\text{or } D \text{ or } C)}$. $\text{MSE}_i^{F(\text{or } D \text{ or } C)}$ is defined as in Eq. (III.10), and calculated for each feature i.

$$\begin{aligned} \text{MSE}_i^F &= \sum_{c \in \mathfrak{C}} (F_i(c) - \bar{F}_i)^2 \\ \bar{F}_i &= \frac{\sum_{c \in \mathfrak{C}} F_i(c)}{|\mathfrak{C}|} \end{aligned} \tag{III.10}$$

where $\mathfrak{C}$ is a set of all the chemical compounds in the dataset. When MSE_i^F is calculated, $\mathfrak{P}$, a set of all the proteins in the dataset is used instead of $\mathfrak{C}$. For G in

Eq. (III.7), the following $\mathrm{MSE}_i^G(t,w)$ is calculated.

$$\mathrm{MSE}_i^G(t,w) = \sum_{c\in\mathfrak{C}} \sum_{j;i+j\in\mathcal{M}(c),i\geq w,I_i\geq t,I_j\geq t} (g_j^c(i) - \bar{g}(i))^2$$

$$\bar{g}(i) = \frac{\sum_{c\in\mathfrak{C}} \sum_{j;i+j\in\mathcal{M}(c),i\geq w,I_i\geq t,I_j\geq t} g_j^c(i)}{\sum_{c\in\mathfrak{C}} \sum_{j;i+j\in\mathcal{M}(c),i\geq w,I_i\geq t,I_j\geq t} 1} \tag{III.11}$$

where t and w are parameters used in equation (III.7). $g_j^c(i)$ is a virtual intensity calculated for a gap between m/z value j and $i+j$ $(i,j>0)$ in a chemical compound c (Eq. (III.7)). $\mathfrak{C}$ is a set of all the chemical compounds in the dataset. $\mathcal{M}(c)$ is a set of all the m/z values observed in a chemical compound c. Here, $g_j(i)$ is used instead of $gap(i)$ as because an intensity for a gap itself is calculated by using $g_j(i)$ in Eq. (III.7).

Then the upper 199 features among ordered features are used for each feature vector type. It is based on the assumption that the larger the MSE for the feature calculated as above is, the more expressive the feature is thought to be. It is assumed that, by evaluating relationship between the more expressive features of different feature vector types, more discriminative relations between different feature vector types can be extracted.

Here, 199 is the number of protein clusters in Eq. (III.5). It is selected because it is independent of the datasets. In addition, it is usually smaller than the number of features for other feature vector types, which increases substantially as chemical compounds in the dataset increase. Many chemical compounds are necessary to realize the comprehensive protein-chemical interaction prediction.

Moreover, in order to equalize the influence of each feature vector type and each feature in the feature types, a normalization scaling was applied. A value for the j-th feature of the sample i was scaled as follows.

$$s(\boldsymbol{x}_{ij}) = -1 + \frac{2(\boldsymbol{x}_{ij} - \min_k \boldsymbol{x}_{kj})}{\max_k \boldsymbol{x}_{kj} - \min_k \boldsymbol{x}_{kj}}$$

2.3.3 One-layer SVM

In contrast to the two-layer SVM model described in the next section, we refer to the prediction model in which pairs of a protein and a chemical compound are mapped to numerical feature space as described in Sec. 2.3.2 and SVM with the kernel function expressed in Eq. (III.9) is applied to the mapped feature space as "one-layer SVM model" (Fig. III.1) in the later sections.

2.3.4 Two-layer SVM

When applied to comprehensive prediction with more than 100,000 prediction targets, "one-layer SVM model" produced many predictions and inevitably yielded many false positives (Table III.9A *random*).

In statistical learning methods including SVM, poor prediction performances frequently result from inadequate training samples. In the previously described SVM

model, random combination of proteins and chemical compounds in datasets except positives are used as negatives. Although combination of two entities generates vast candidates for negatives, we can use only a limited amount of them due to limitation of time and computation in both learning and feasible comprehensive application. Moreover, imbalance between the number of positive samples and that of negatives frequently leads to improper classification models. Therefore, it is assumed that predicted false positives lie in the outside of the coverage of the classification rule extracted from a limited region of protein-chemical interaction space.

In order to extend the coverage of proper prediction, in the two-layer SVM, we utilize several "one-layer SVM models" each of which embodies a classification rule derived from a different dataset with different negatives and covers a different region of protein-chemical interaction space.

Since there existed no huge mass spectra databases containing more than one million chemical compounds and suitable for comprehensive binding ligand prediction as of 2008, the two-layer SVM utilized, along with protein sequence data, only chemical structure data to represent chemical compounds and the DrugBank2 dataset to construct prediction models.

■First-layer SVM

Based on feature vectors for a pair of a protein and a chemical compound respectively expressed in Eqs. (III.5) and (III.8) and similarity between two samples defined in the following modified Eq.(III.9),

$$K\{(C,D),(C',D')\} = \prod_{I\in\{C,D\},J\in\{C',D'\}} k_{IJ(=JI)}(I,J), \quad k_{IJ}(\boldsymbol{x},\boldsymbol{y}) = \exp(-\gamma_{IJ}\|\boldsymbol{x}-\boldsymbol{y}\|^2), \tag{III.12}$$

we generated 100 first-layer SVM models with different random combination of proteins and chemical compounds as negatives (Fig. III.1). The SVM parameters were chosen to give the best accuracy in 10-fold cross validation in one set of positives and negatives.

In this study, we prepared two sets of first-layer SVM models, each of which consists of 100 models. One set *allpos* contains the SVM models constructed from 1,731 positives, or the whole DrugBank2 dataset (*allpos* first-layer SVM models where D_0^l in Eq. (III.8) was used), and 1,750 negatives. The other set *subpos* is composed of models with 534 positives, one of 10 kinds of DrugBank2 subsets, and 550 negatives (*subpos* first-layer SVM models where D_2^l in Eq. (III.8) was used). A protein found n times in the DrugBank2 dataset is designed to appear $\lfloor n/10 \rfloor + 1$ times in a DrugBank2 subset, and the chemical compounds with which the protein forms a pair differ between different subsets.

■Second-layer SVM

The second-layer SVM directly utilizes the outputs of the first-layer SVM models as inputs. The second-layer SVM model was constructed from the whole DrugBank2 dataset and reasonably designed negatives, which are described in detail later in Sec. 2.4 on the basis of the RBF kernel $K(\boldsymbol{x},\boldsymbol{y}) = \exp(-\gamma\|\boldsymbol{x}-\boldsymbol{y}\|^2)$ in the Eq. (III.1). The SVM parameters were selected in such a way that they gave the best

accuracy in the 10-fold cross validation. Finally, the two-layer SVM functions as shown in Fig. III.2.

For comparison of classification methods, quadratic discriminant analysis (QDA) and artificial neural network (ANN), which were implemented by R package *qda* and *nnet* (`http://cran.r-project.org/`), were applied to outputs of the first-layer SVM models.

■Feature selection

The number of the first-layer SVM models whose output is used in the second-layer SVM models mainly determines the computation time and the workload of the two-layer SVM methods. Therefore, in order to realize comprehensive protein-chemical interaction predictions, fewer first-layer models achieving the high prediction accuracy are given preference.

We applied he recursive feature elimination (RFE) method (Xue *et al.*, 2004) in order to determine the first-layer SVM models used to construct the second-layer SVM model. When n (=100 for the first time) first-layer models are considered, the model i satisfying the following criterion is eliminated to produce the second-layer SVM model with $n-1$ dimensions.

$$\underset{i}{\operatorname{argmin}} \ \frac{1}{2}\boldsymbol{\alpha}^t H(0)\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^t H(i)\boldsymbol{\alpha},\ H(i) = \begin{bmatrix} K'_{11} & \cdots & K'_{1n} \\ \vdots & \ddots & \vdots \\ K'_{n1} & \cdots & K'_{nn} \end{bmatrix},\ K'_{jk} = y_j y_k \exp\left(-\gamma \sum_{l=1, l\neq i}^{m} (x_{jl} - x_{kl})^2\right), \tag{III.13}$$

where $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{\alpha}$ are the same as those in Eq. (III.1). This elimination continues until n reaches a certain number.

2.4 Negative data design

As described in Sec. 2.3.4, the number of false positives, one of the common problems of computational prediction approaches, can be reduced by carefully designing training samples.

We followed and modified the method described in Ref. Wang *et al.*, 2006 for the design of negative data leading to the reduction of the number of false positives.

From P:(positive samples or the DrugBank2 dataset), U:(all the possible combinations of proteins and chemical compounds found in the DrugBank2 dataset except positive samples), $N = \emptyset$, we constructed the designed negative dataset N through following two phases.

(Phase A) Determination of negative dataset seed.

1. Add a sample i satisfying the following criterion to N.

$$\max_{\boldsymbol{x}_i \in U} d(\boldsymbol{x}_i, P), \quad d(\boldsymbol{x}_i, P) = \min_{\boldsymbol{x}_j \in P} \|\boldsymbol{x}_i, \boldsymbol{x}_j\|$$

2. Add a sample i satisfying the following criterion to N.

$$\max_{\boldsymbol{x}_i \in (U-N)} \left[d(\boldsymbol{x}_i, P) \cdot \sum_{\boldsymbol{x}_j \in N} d(\boldsymbol{x}_i, \boldsymbol{x}_j) \right], \quad d(\boldsymbol{x}_i, \boldsymbol{x}_j) = \|\boldsymbol{x}_i, \boldsymbol{x}_j\|$$

3. Repeat the step 2 until $|N|$ reaches a certain number ($|N| = 3500 \approx 2 \times |P|$).

(Phase B) Expansion of negative dataset.

1. Construct a SVM model from P and N.
2. The constructed SVM model is applied to $U-N$. L samples ($|L| = 3500 \approx 2\times|P|$) are added to N according to the probabilistic output p_i of SVM expressed in Eq. (III.2) and following rules.

min : Top L samples in the ascending order of $p_i,\ i \in (U-N)$ s.t. $p_i \leq 1$.
max : Top L samples in the descending order of $p_i,\ i \in (U-N)$ s.t. $p_i \leq 1$.
mle : Top L samples in the descending order of $p_i,\ i \in (U-N)$ s.t. $p_i \leq 0.5$.
mlt : Top L samples in the descending order of $p_i,\ i \in (U-N)$ s.t. $p_i < 0.5$.

3. Repeat the step 1 and 2.

Particularly, the rules *max*, *mle* and *mlt* were introduced for significant false positive reduction.

2.5 Strategy of feedback and supplement with additional data

Computational predictions by statistical methods , docking methods or molecular dynamics methods involve success and failure after they are verified by biological experiments. One of the merits of using statistical methods involving training with known data is that results obtained by verification experiments can be efficiently utilized or feedbacked to produce newer and more reliable predictions. Moreover, without verification experiments, additional data can be acquired from, for example, literature and used to enhance prediction reliability.

Given N_p positive and N_n negative samples in known data and M_p positives and M_n negatives in additional or feedback data, a straightforward strategy for the integration of additional data in statistical training such as SVM is to train a statistical model based on a dataset consisting of $N_p + M_p$ positives and $N_n + M_n$ negatives. When the two-layer SVM strategy is considered, another strategy of feedback and supplement involves the utilization of an additional model based on additional data. In this strategy, the second-layer SVM is trained on the basis of $N_p + M_p$ positives and $N_n + M_n$ negatives, and a sample s_i in the second-layer is represented as follows,

$$s_i = (w \times p_i^a, p_i^1, \ldots, p_i^k).$$

Here, p_i^a is an output of the additional model trained on the basis of M_p positives and M_n negatives. p_i^j is an output of the first-layer SVM model j and w is a weighting factor.

For three proteins; UniProt ID P10275 (androgen receptor), P11299 (muscarinic acetylcholine receptor M1) and P353367 (histamine H1 receptor), ligand data that was not included in the DrugBank2 dataset were collected from literature (Funder and Mercer, 1979; Link *et al.*, 2005; Kinoyama *et al.*, 2006) and public databases; PDSP Ki database (Roth *et al.*, 2000) and GLIDA (Okuno *et al.*, 2008) in February 2008. Overall, 35 androgen receptor-ligand pairs, 49 muscarinic acetylcholine receptor M1-ligand pairs and 1060 histamine H1 receptor-ligand pairs were supplemented. Additional models were constructed using these supplemental pairs as positives and

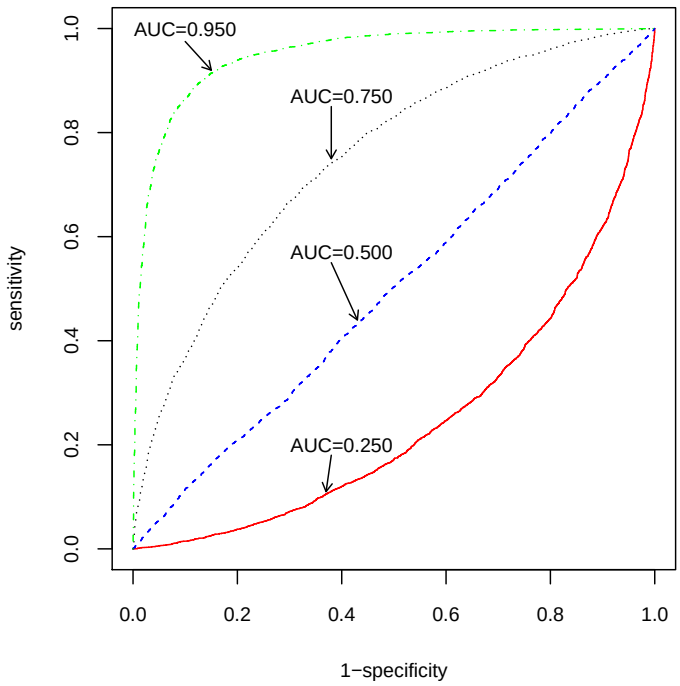


Fig. III.7 Example of ROC curves. ROC curves of different AUC values were derived from 10,000 samples.

regarding pairs of each protein and chemical compounds in the DrugBank2 dataset other than binding ligands and those in negative dataset seed described in Sec. 2.4, or pairs least likely to interact, as negatives. Roughly the same number of these two types of negatives were utilized. When these supplemental pairs had been regarded as negatives in the process of selection of candidates for negatives, these samples were treated as positives.

2.6 Analyses of predictions

2.6.1 Evaluation of the prediction performances

In experiments using previously described datasets, we evaluated the prediction performances of our method using the 10-fold cross-validation on the basis of several measurements including precision (prec.), sensitivity (sens.), specificity (spec.), accuracy (acc.), and Matthew's correlation coefficient (MCC), and area under the ROC curve (AUC). These measurements are expressed as follows.

$$\text{prec.} = \frac{TP}{TP+FP},\ \text{sens.} = \frac{TP}{TP+FN},\ \text{spec.} = \frac{TN}{TN+FP},\ \text{acc.} = \frac{TP+TN}{TP+FP+TN+FN}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FN) \times (TN+FP) \times (TP+FP) \times (TN+FN)}}$$

	Model 1		Model 2	
(protein)	Compound α	Compound β	Compound α	Compound β
a	◯	◯	×	×
b	◯	◯	◯	◯
c	◯	×	◯	×
d	×	◯	×	◯
e	◯	×	◯	×

Fig. III.8 Similarity between compound α and β.

TP: true positives or a number of known positives predicted as positive. FP: false positives, or a number of negatives predicted as positive. FN: false negatives, or a number of known positives predicted as negative. TN: true negatives, or a number of negatives predicted as negative.

AUC is based on the ROC curve which plots "1−specificity" (=x axis) and "sensitivity" (=y axis) at all the possible thresholds (Fig. III.7).

In comprehensive binding prediction, we utilized the two measurements, rec_x and "evaluation". rec_x is the recall rate ($= TP/(TP+FN)$) at the threshold of x, ranging from 0 to 1. $x = 0.5$ is the threshold following the definition of SVM. "evaluation" was calculated as follows.

$$\begin{aligned}\text{evaluation} = \quad & 100 \times \left(\frac{1}{2} \left[\mathrm{rec}_{0.5} + \frac{\mathrm{rec}_{0.95} + \mathrm{prec}_{0.95}}{2\{1 + (1 - \mathrm{rec}_{0.95})(1 - \mathrm{prec}_{0.95})\}} \right] \right. \\ & \left. - \frac{\text{total \# of predicted positives - \# of known positives}}{\text{total \# of prediction targets - \# of known positives}} \right)\end{aligned} \tag{III.14}$$

Here, prec_x is the precision at the threshold of x.

This measure "evaluation" is based on the following three ideas. First, only prediction results beyond some high threshold are frequently considered. Secondly, the prediction of 20% sensitivity and 80% precision is often preferred to that of 50% sensitivity and 50% precision. Thirdly, prediction with not too many candidates and high sensitivity at some lower threshold can be a target for comprehensively applicable experimental methods.

For all of these measurements, the higher the value, the better the prediction is.

2.6.2 Similarity measure based on predicted proteins

Target protein sets generated by comprehensive target protein predictions for chemical compounds can reveal biological and functional similarities among these chemical compounds.

It is generally assumed that the more target proteins two drugs have in common, the more biologically or functionally similar they are thought to be as most drugs function via binding proteins.

Therefore, in this study, we define the similarity between two chemical compounds

α and β as follows,

$$s(\alpha,\beta) = \begin{cases} \dfrac{1}{n^2} \displaystyle\sum_{i\in\{1,\ldots,n\}} \sum_{j\in\{1,\ldots,n\}} \dfrac{2\times r(A_i,B_j)}{r(A_i,A_j)+r(B_i,B_j)}, \\ \qquad\qquad \text{if } \forall i,j \;\; r(A_i,B_j), r(A_i,A_j), r(B_i,B_j) \neq 0 \\ 0 \qquad\qquad \text{otherwise} \end{cases} \tag{III.15}$$

$$r(A,B) = \frac{|A\cap B|}{|A\cup B|}$$

where A_i is a target protein set predicted by an SVM model i for a chemical compound α. Here, to overcome the problem of most statistical learning methods that they depend on limited training data, several prediction results made by models with different negative samples are used for the sake of higher confidence.

The higher the $s(\alpha,\beta)$ value, the more biologically similar α and β are thought to be.

For example, when compound α and β have target proteins a, b, c, d and e (Fig. III.8), $r(A_1,\ B_1)$ in Eq. (III.15) is represented as follows with target proteins of the compound α predicted by the model 1, $A_1 = \{a,\ b,\ c,\ e\}$ and $B_1 = \{a,\ b,\ d\}$.

$$r(A_1,B_1) = \frac{A_1\cap B_1}{A_1\cup B_1} = \frac{|\{a,b\}|}{|\{a,b,c,d,e\}|} = \frac{2}{5} = 0.4.$$

$s(\alpha,\beta)$ can be calculated as follows,

$$\begin{aligned} s(\alpha,\beta) = \;& \frac{1}{4}\sum_{i\in\{1,2\}}\sum_{j\in\{1,2\}} \frac{2\times r(A_i,B_j)}{r(A_i,A_j)+r(B_i,B_j)} \\ = \;& \frac{1}{4}\Big(\frac{2\times r(A_1,B_1)}{r(A_1,A_1)+r(B_1,B_1)} + \frac{2\times r(A_1,B_2)}{r(A_1,A_2)+r(B_1,B_2)} + \\ & \frac{2\times r(A_2,B_1)}{r(A_2,A_1)+r(B_2,B_1)} + \frac{2\times r(A_2,B_2)}{r(A_2,A_2)+r(B_2,B_2)}\Big) \\ = \;& \frac{1}{4}\left(0.4 + \frac{2\times 0.2}{0.75+2/3} + \frac{2\times 0.2}{0.75+2/3} + 0.25\right) \\ = \;& 0.304. \end{aligned}$$

In Fig. III.8, the protein a is predicted as target for both the compound α and β by the model 1. It may be true, but the fact that the protein a is not predicted by the model 2 suggests that it derive from defects of the model 1. It can be the case that the model 1 has the protein a in positive protein-chemical pairs, but not in negative samples. In such a case, the model 1 may tend to predict a pair of the protein a and a chemical compound as positive. Eq. (III.15) aims to cope with such cases by comparing prediction results of different models.

Principal component analysis (PCA) was then applied to the similarity matrix S, whose element s_{ij} represents the similarity between the compounds i and j.

Wet experiment section

2.7 Target protein - human androgen receptor

Androgen receptor (AR) is one of genes responsible for prostate cancer, which is the most frequently diagnosed cancer in men in the United States according to the American Cancer Society Statistics for 2008. AR is a steroid hormone receptor and a transcription factor belonging to the nuclear receptor superfamily. AR protein consists of the N-terminal domain that contains the activation function 1 region and regulates the transcription activity (Danielian *et al.*, 1992), the DNA binding domain at the central, the ligand binding domain at the C-termini, and the hinge region containing nuclear localization signals between these binding domains. In prostate cancer, mutations in the AR gene, overexpression of the AR proteins, and further suppression of cancer cell proliferation by knockdown of the AR gene induced by siRNA were observed (Compagno *et al.*, 2007).

In prostate cancer therapy, hormone therapy using an androgen antagonist such as flutamide, nilutamide and bicalutamide exists. These drugs are indicated to cause severe side effects such as interstitial pneumonia and including liver disorders as AR is expressed in several tissues including the lungs and liver. However, a selective androgen receptor modulator, which acts as an antagonist in specific tissues and as an agonist in other tissues or vice versa, is expected to overcome side effects (Gao *et al.*, 2005).

As a mechanism of action of selective androgen receptor modulators is not fully elucidated and most of them have been found by chance (Chen *et al.*, 2002), it is necessary to efficiently identify a lot of compounds targeting AR and select potent antagonists of prostate cancer cells from them for the discovery of selective androgen receptor modulator drugs.

2.8 Materials

Unless otherwise specified, all solvents and reagents were obtained from commercial suppliers.

In the plasmid preparation, pTriAR, a construct in which Androgen receptor (AR) cDNA is subcloned into the pTriEX-3 Neo vector, was provided by Taiho Pharmaceutical Co., Ltd.

In the *in vitro* binding assay, dihydrotestosterone (DHT), flutamide, nilutamide, spironolactone and cortexolone were purchased from Sigma. Testosterone and bicalutamide were purchased from Wako Pure Chemical Industries, Ltd. ZINC 04369595, MDPI 944, MDPI 1011, NSC 6129, MDPI 10314, 3-epiuzarigenin, ZINC 04026296, methandriol, vitamin D3, ZINC 03849821, P712100 and fluanisone were purchased from Namiki Shoji Co., Ltd.

2.9 Plasmid preparation

The gene sequences corresponding to the ligand-binding domain (609th a.a. - 919th a.a.) of androgen receptor C termini (ARC) were amplified by PCR with pTriAR as a template, 5′-ATGACTCTGGGAGCCCGG-3′ (sense) and 5′-CCCTCGAGTCACTGGGTGTGGAAATAGATGGG-3′ (anti-sense) primers, and KOD plus (Toyobo Co., Ltd.) as DNA polymerase. The PCR conditions were as follows: 40 cycles of denaturation (98 ℃, 15 seconds), annealing (60 ℃, 30 seconds) and extension (68 ℃, 3 minutes).

After agarose gel electrophoresis of PCR products, DNA fragments of supposed ARC were subcloned into pBlueScript II SK(-) vector (Stratagene) with *EcoRV*. This recombinant plasmid (pBS/ARC) was sequenced to verify that ARC was properly amplified.

After verification, ARC sequences were digested from pBS/ARC, and subcloned into pMALc-2x vector digested with *HindIII* and *BamHI* to obtain a recombinant plasmid pMAL/ARC which expressed, in *E. coli*, the maltose binding protein tagged androgen receptor C-termini (MBP-ARC).

Here, it is reported that an *in vitro* binding assay with ARC produced almost the same result as that with the whole length AR (Zhu *et al.*, 2001).

2.10 Recombinant ARC protein preparation and purification

The pMAL/ARC plasmid was transfected into *E. coli* DH5α. The transfected cells were cultivated in LB medium containing 50 μg/ml ampicillin at 37 ℃ overnight. The culture solution was diluted by LB medium so that OD_{600} was equal to 0.1. After one hour cultivation of the diluted culture solution at 25 ℃, 0.1 mM IPTG was added to the solution. Then, the solution was cultivated at 25 ℃ for 12 hours.

After cultivation, the cells were harvested with centrifuge at 3,500 rpm for 30 minutes. The cells were then suspended in 30-ml MT-PBS containing 1 % Triton and disintegrated with an ultrasonic homogenizer.

After 20-minute centrifugation of the homogenized solution at 3,500 rpm, a supernatant was collected. The supernatant was further centrifuged at 13,000 rpm for 30 minutes to obtain a soluble fraction. The soluble fraction was applied to a 10-ml column of amylose resin (New England BioLabs), which was washed with a 50-ml column buffer consisting of 20 mM Tris-Hcol pH 7.4, 200 mM NaCl and 1mM EDTA. After column washing with the 125-ml column buffer containing 0.03 mM maltose, the recombinant ARC protein was eluted with 15-ml column buffer containing 10mM maltose. All the purification processes were carried out at 4 ℃.

2.11 The *in vitro* binding assay - hydroxyapatite method

50 μg/ml recombinant ARC protein , 2 nM [^{3}H]-DHT and a test compound in a molar ratio x:1 ($x = 0.01 - 3 \times 10^5$) with [^{3}H]-DHT were mixed in a binding buffer

consisting of 50mM Tris-HCl pH 7.4, 800 mM NaCl, 10 % glycerol, 1mg/ml BSA and 2mM DTT to obtain a 100 μl mixture solution. The mixture solution was incubated at 4 ℃for 3 hours.

After incubation, BioGel HT (Bio-Rad Laboratories) was added to the solution and incubated on ice for 15 minutes, during which vortexing was done every 5 minutes. The incubated solution was centrifuged at 1,000 rpm for a minute and a supernatant was removed.

A wash buffer consisting of 40 mM Tris-HCl pH 7.6, 100 mM KCl, 1mM EDTA and 1mM EGTA and cooled on ice was added to the precipitate by 1 ml and mixed. The produced solution was centrifuged at 1,000 rpm and a supernatant was removed. This process was repeated three times.

The collected precipitate was suspended in 3 ml Aquasol (New England nuclear Corp.). Radioactivity of the suspended solution in a scintillation vial (Perkin Elmer) was measured by by a liquid scintillation counter.

Here, the radioactivity derived from [^{3}H]-DHT showed the amount of the recombinant ARC protein bound by [^{3}H]-DHT. As [^{3}H]-DHT and test compounds competitively were thought to bind to the recombinant ARC protein, decrease of radioactivity from that measured without competitors, or when [^{3}H]-DHT were thought to bind to all the recombinant ARC protein, reflected the content of the recombinant ARC protein bound by unlabeled competitors. Therefore, the concentration of the test compound to [^{3}H]-DHT in which the measured radioactivity corresponded to 50 % of that measured without the test compounds was regarded as IC_{50} of the test compound.

3 Results

3.1 Proof of applicability of statistical learning methods by using one-layer SVM

3.1.1 Specific binding prediction

■Evaluation of one-layer SVM

We define "specific binding prediction problem" as the prediction of all possible interactions between the chemical compounds being tested and a specific family of proteins. We compare and contrast this with "general binding prediction" at a later stage in the text. It has often been observed that compounds designed against one protein target also demonstrate useful activities against other members of the same protein family. This suggests that the members of a particular protein family may often share a common essential binding mechanism. The aim of our specific binding model is to elucidate this shared mechanism and exploit it in the classification of protein-chemical pairs as binding and non-binding.

In our computational assessments of specific binding predictions, a prediction model for the human adrenergic receptor family was constructed from the ADR drugs dataset. The prediction performance of this model was the evaluated using a 10-

Table III.3 Prediction performances in the ADR drug dataset

[1]vector type		pre.(%)	sen.(%)	acc.(%)	MCC
(A) Specific binding prediction					
[2]$(C, F, G^{12}_{15})^{c}_{\mathrm{LR}}$		65.7	50.0	75.0	0.404
[3]$(C, F, G^{12}_{15})^{c}_{\mathrm{NN}}$		60.9	76.7	76.2	0.502
[4]$(C, F, G^{12}_{15})^{c}_{\mathrm{lin.}}$		73.0	51.4	77.8	0.469
[5]$(C, F, G^{12}_{15})^{c}_{\mathrm{rbf}}$		88.3	79.6	89.8	0.765
[6]$(C, F, G^{12}_{15})^{c}_{\mathrm{di}}$		85.6	75.4	87.7	0.716
$(C, F, G^{12}_{15})^{k}$		89.7	85.9	92.1	0.820
(C, F)		88.1	83.8	91.0	0.793
$(C, G^{12}_{7.5})$		83.2	76.8	87.3	0.707
(C, D^{8}_{0})		90.4	93.0	94.4	0.875
$(C, D^{8}_{2}, F, G^{12}_{7.5})$		93.5	91.5	95.1	0.889
(B) Classification of agonism and antagonism					
[7](C, F, G^{12}_{0}) (=pair)		98.6	100.0	99.3	0.986
[7](F, G^{12}_{15}) (=compound)		88.5	88.5	87.5	0.748
(C) Prediction based on different regions of proteins					
vector type	[8]region	pre.(%)	sen.(%)	acc.(%)	MCC
$(C, F, G^{12}_{7.5})$	TMH	90.6	88.7	93.3	0.847
$(C, F, G^{12}_{7.5})$	EL	82.6	80.3	88.0	0.725
(C, F, G^{12}_{15})	CL	80.1	79.6	86.8	0.700

[1] "c" means that concatenation of feature vectors was used for combination of vectors to represent a sample at step 3 in Fig III.2. "k", on the other hand, means that combination of kernels in equation (III.9) was exploited. If not specified, "k" was applied. [2] The logistic regression was applied (R package brlr (Firth, 1993) was used). [3] The artificial neural network was applied (R package nnet (Ripley, 1996) was used). [4] The SVM with linear kernel was applied. [5] The SVM with RBF kernel was applied. [6] Dipeptide composition was used in mapping C and the SVM with RBF kernel was applied. [7] If mapping C was used or a pair is considered, 142 protein-chemical pairs were treated. If not, 48 compounds only were considered. [8] TMH transmembrane helix; El extracellular loop; CL cytoplasmic loop. The sequences of each region were used to represent the feature vector C.

fold cross validation and some prediction performance measurements (Table III.3A).

Two main features of our proposed method is the representative description of proteins and compounds and the representation of a protein-chemical pair. In our current study, we proposed the representation of a protein-chemical pair by multiplication of several kernel functions (equation (III.9)). This type of representation gave a better performance (0.820 MCC) than just concatenating feature vectors to represent a pair (0.765 MCC) (Table III.3A). This result indicates the importance of considering the crossover effects between different types of feature vectors.

Table III.3A also shows the validity of using non-linear SVM for the classification of binding and non-binding protein-chemical pairs. As shown in Table III.3A, SVM using the RBF kernel showed the best accuracy (89.8%) when the same combination of feature vectors and the same way of representing a pair (concatenation of vectors) was used (Table III.7A). The logistic regression, the artificial neural network and SVM with the linear kernel gave the same level of prediction performances (75-78%

Table III.4 Effects of using the intensities of mass spectra in classification of the ADR dataset

vector type	binding expression	prec.(%)	sens.(%)	acc.(%)	MCC	AUC
(C, F, G_{15}^{12})	[1]conc.	88.3	79.6	89.8	0.765	0.941
(C, F)	conc.	78.3	83.8	87.0	0.712	0.914
(C, G_{15}^{12})	conc.	81.3	70.4	85.0	0.650	0.885
$(C,^2 F',^3 G_{15}^{12'})$	conc.	82.0	73.9	86.1	0.679	0.889
(C, F', G_{15}^{12})	conc.	82.9	71.8	85.9	0.672	0.920
$(C, F, G_{15}^{12'})$	conc.	83.2	76.8	87.3	0.707	0.918
(C, F')	conc.	81.4	73.9	85.9	0.674	0.909
$(C, G_{15}^{12'})$	conc.	80.8	71.1	85.0	0.651	0.875

[1]: Using concatenation to express binding. A simpler way to express interaction than that in Methods. For two sample vectors S_1 and S_2, which consisted of several types of feature vectors (for example, $S_1 = (C_1, F_1, G_1)$), they are regarded as one vector respectively. The similarity between them used in equation (5) in Methods is calculated as follows.

$$K_{\mathrm{conc}}(S_1, S_2) \equiv \exp(-\gamma \|S_1 - S_2\|^2)$$

Each feature vector type is allowed to have a different vector length or the number of features.
[2]: F' doesn't consider the intensity and is calculated as follows like equation (2) in Methods.

$$F'(c) = (\phi_c(m))_{m \in \mathcal{M}}, \quad \phi_c(m) = \begin{cases} 1 & \text{if } m \in \mathcal{M}(c) \\ 0 & \text{otherwise} \end{cases}$$

[3]: G' doesn't consider the intensity and is calculated as follows.

$$G_t^{w'}(c) = (\xi_c(m))_{m \in \mathcal{M}_g}, \quad \xi_c(m) = \begin{cases} 1 & \text{if } m \in \mathcal{M}_g(c) \\ 0 & \text{otherwise} \end{cases}$$

accuracy) (Table III.7A).

We introduced four types of feature mappings; C for protein description, and D, F and G for representing chemical compounds. Mapping C is derived from the frequency of subsequences and physico-chemical properties of amino acids. As shown in Table III.7A, this feature mapping of proteins showed a better performance (0.765 MCC) than the commonly used dipeptide frequency (0.716 MCC) with fewer features (199 vs. 400).

For chemical compound description, mapping D is based on chemical structure data, and both F and G are derived from mass spectrometry data. The use of D gave very high prediction performances such as 94.4% accuracy (Table III.3A). On the other hand, the combination of F and G achieved a bit lower than the use of D, but significantly high performances, including a 92.1% prediction accuracy, and a more than 0.8 MCC (Table III.3A). Moreover, the combination of D, F and G showed the best performances in Table III.3A, including 0.889 MCC.

The three mapping D, F and G are based on a common principle that extracted substructures of chemical compounds are sufficiently representative of that compound that they can be used to elucidate the binding mechanism. Though mass spectra are more unprocessed data than chemical structures, the peaks in the mass spectra for F and G can be interpreted as substructures, and the results show that it works

(A) ADR drug dataset (B) DrugBank dataset

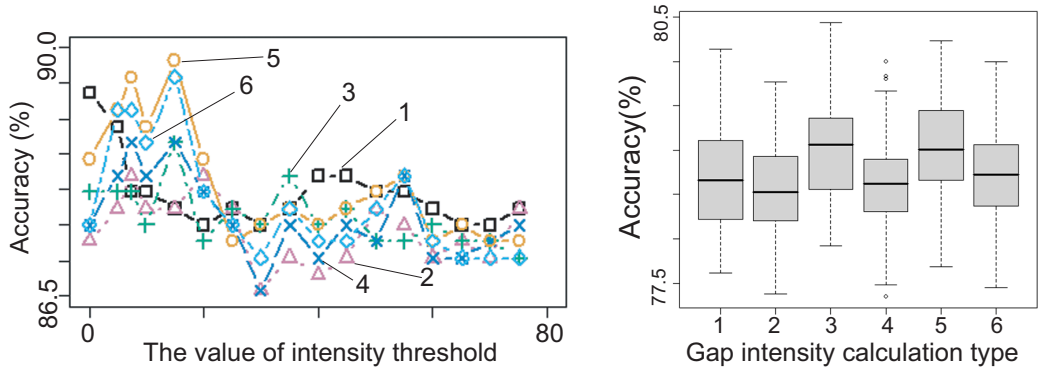


Fig. III.9 The prediction results of different gap intensity calculations. The numbering is based on Eq. (III.16). Three types of vectors C, F and G_t^{12} in Eq. (III.5), (III.6) and (III.16) were concatenated to represent a sample. (A) the prediction results on the ADR drug dataset (see Sec. 2.2) with different threshold t in Eq. (III.16). (B) the prediction results on the DrugBank datasets (see Sec. 2.2) with $t = 0$. As the DrugBank dataset treated random protein-chemical pairs as negative samples, evaluation was conducted on 100 DrugBank datasets with different negative samples

sufficiently.

In comparison with D, one possible disadvantage of using a combination of F and G is the existence of synonyms, or compounds whose chemical structures are different but whose molecular weights, or m/z values in the spectra, are equivalent. This is also thought to be the reason why G showed a lower performance (0.707 MCC) than F (0.793 MCC). On the other hand, one advantage of using the mapping method based on mass spectra is the existence of intensities that reflect the physical-chemical properties of each peak. As shown in Table III.4, the use of intensity generally improved the prediction performances.

As for the virtual intensity for gaps, as shown in Fig III.9, the gap intensity calculation function used in Eq. (III.7), or 5 in Fig. III.9, performed generally well among the following 6 functions.

$$
\begin{aligned}
&1.\ g_i(j-i) = \frac{I_i + I_j}{2} && 2.\ g_i(j-i) = I_i \times \frac{1/I_j}{\sum_{k;k>i,I_k \geq t} 1/I_k} \\
&3.\ g_i(j-i) = I_i \times \frac{I_j}{\sum_{k;k>i,I_k \geq t} I_k} && 4^*.\ g_i(j-i) = I_i \times \frac{-\ln(I_j)}{\sum_{k;k>i,I_k \geq t} -\ln(I_k)} \\
&5.\ g_i(j-i) = I_i \times \frac{\ln(I_j)}{\sum_{k;k>i,I_k \geq t} \ln(I_k)} && 6^*.\ g_i(j-i) = I_i \times \frac{\exp(I_j)}{\sum_{k;k>i,I_k \geq t} \exp(I_k)}
\end{aligned}
\qquad \text{(III.16)}
$$

$^*\ I_k \in (0,1)$

Hence, based upon these performances assessments, the integration of D, F and G mapping has the capacity to compensate for the limitations that are inherent in each

individual mapping method and thus produce more accurate predictions (Table III.3A and III.5).

Overall, the best result found in these analyses was a 95.1% accuracy (Table III.3A). These very high values indicate that an essential binding mechanism shared among protein family members can be extracted statistically by SVM from a large data set that contains adequate feature vectors for protein-chemical pairs.

■Prediction of binding properties: classifications of agonism and antagonism

In our current study, we represented a sample by combining feature vectors for proteins and chemical compounds, and classify protein-chemical pairs to predict interactions between them. To show the effectiveness of this representation, we conducted the following experiment.

The ADR drug dataset comprises 22 agonists constituting 73 receptor-agonist pairs, and 26 antagonists for 69 receptor-antagonist pairs. To predict whether a compound acts as an agonist or an antagonist, two types of classification tasks were performed. The first of these is a classification of agonist-receptor pair and antagonist-receptor pair in which a protein-chemical pair is the input, and the second is a classification of agonist and antagonist where only the chemical compounds are used as the input.

The results of this analysis are shown in Table III.3B, and indicate that, for the prediction of either agonism or antagonism of the adrenergic receptor by different chemical compounds, our classification of protein-chemical pairs gave a better performance (0.986 MCC) than classification of chemical compounds alone (0.748 MCC). These findings suggest the usefulness of considering protein-chemical pairs.

Table III.3B also suggests that some activating and non-activating binding mechanisms can be extracted from the feature vectors of protein-chemical pairs by SVM. Moreover, this method may be applied also to the prediction of other binding properties such as affinity, where samples are classified into two classes by fixed threshold or regression methods such as support vector regression.

■Predictions based on different regions of proteins

An adrenergic receptor, which is also a G-protein coupled receptor (GPCR), consists of three regions; TMHs (transmembrane helices), ELs (extracellular loops) and CLs (cytoplasmic loops). Moreover, the majority of the small-molecule drugs that have been developed interact with the seven transmembrane-spanning domains of GPCRs (Kristiansen, 2004). In our computational analysis of the ADR drug binding predictions using each region of the GPCRs, the utilization of TMHs alone in a C mapping gave a better performance (93.3% accuracy) than that of the whole sequence (92.1% accuracy), EL (88.0% accuracy) or CL (86.8% accuracy) (Table III.3C).

This result may indicate the biological relevance of this protein-chemical interaction prediction. In addition, it suggests the possibility that our novel prediction method can successfully identify protein regions that are essential for the binding of small molecules.

Table III.5 General prediction performance in the DrugBank dataset

vector type	prec.(%)	sens.(%)	acc.(%)	MCC
$(C, F, G_0^{12})^{1*}$	76.2 ± 0.2	71.6 ± 0.2	74.9 ± 0.2	0.498 ± 0.003
$(C, F, G_0^{12})^2$	75.8 ± 0.2	60.6 ± 0.2	80.7 ± 0.1	0.546 ± 0.002
$(C, F, G_0^{12})^3$	75.1 ± 0.2	54.0 ± 0.2	84.3 ± 0.1	0.544 ± 0.002
$(C, D_0^6)^2$	81.9 ± 0.2	66.5 ± 0.2	84.2 ± 0.1	0.630 ± 0.002
$(C, D_0^6, F, G_{7.5}^{12})^2$	84.6 ± 0.2	64.1 ± 0.2	84.4 ± 0.1	0.634 ± 0.002

* shows a number of random pairs generated to produce negative samples for constructing SVM models. For each number, 100 different negative sets were generated and evaluated. [1], [2] and [3] mean 1,000, 2,000 and 3,000 random pairs respectively.

3.1.2 General binding prediction

We define "general binding prediction problem" as the prediction of the interactions between chemical compounds and proteins belonging to different protein families. Hence, our general binding prediction model is designed to extract some of the underlying common binding mechanisms that are shared by several binding protein families and utilizes this for general protein-chemical interaction predictions.

In our computational experiments for general binding predictions, the general binding models were constructed from the DrugBank dataset. The prediction performances for different negative samples within this model were evaluated as shown in Table III.5. This method achieved more than 80% accuracy for most negative sample numbers (Table III.5). Based upon this relatively high performance, we conclude that some general binding mechanisms that are common to a number of protein families can be successfully detected by our proposed method and that its application enables us much wider series of predictions.

Though we used random pairs of drugs and proteins as negative samples in constructing a model, the lack of reliable negative samples is always a problem when applying the statistical learning methods. In our current study, it is assumed that drugs in the DrugBank dataset rarely interact with proteins other than their known targets because they are approved drugs. Moreover, to see the tolerance of our method to accidentally containing positive drug-protein pair in a negative sample set, we conducted an experiment in which a fraction of positive samples were intentionally labeled as negatives (pseudo-negatives). We successfully observed that those pseudo-negatives were predicted as positives until the number of pseudo-negatives exceeded a certain level (Fig. III.10)). Hence, our proposed method is robust to a small fraction of unknown positives in negatives which may be the case in using approved drugs.

3.1.3 Indication of biological validity of statistical approaches

In bioinformatics, statistical approaches extract rules from numerical data corresponding to biological properties. Here, it is not guaranteed that the extracted rules are biologically valid, and further some general rules may be obtained statistically from any kind of numerical data which are meaningless and irrelevant to biological

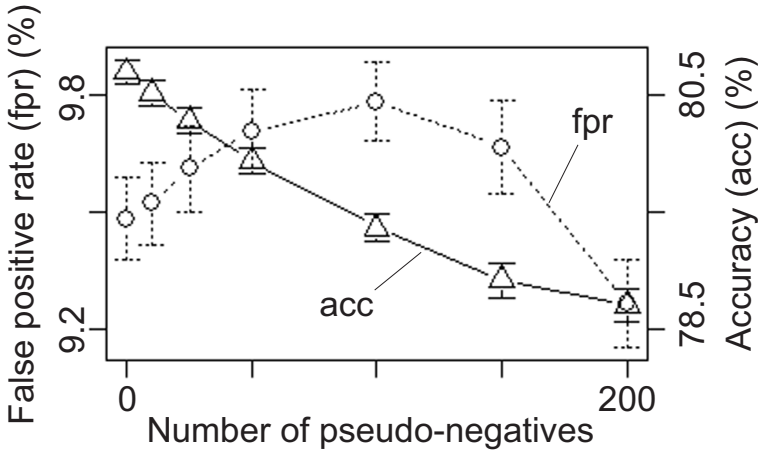


Fig. III.10 The relationship between false positive rate or accuracy and a number of pseudo-negatives. False positive rate (fpr) = $FP/(TN + FP)$. A pseudo-negative is a positive sample that is intentionally treated as negative.

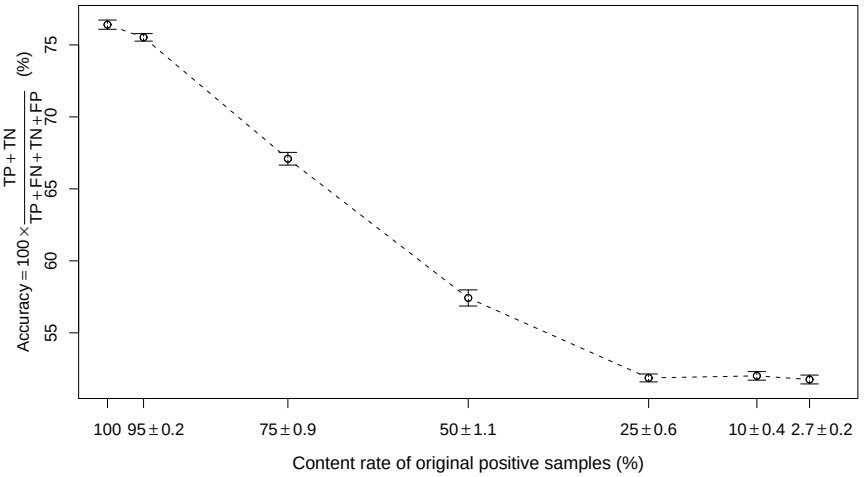


Fig. III.11 The prediction accuracy is proportional to the content rate of biologically valid samples. The average of 10 datasets produced by shuffling pairs corresponded to each content rate (ex. 50%) of pairs comprising a protein and a chemical compound in the original dataset. A usual SVM training referred to as one-layer SVM in Methods and a 10-fold cross-validation evaluation were performed for each dataset of 1,731 positives and 1,750 negatives (or random pairs other than positives). Here, the SVM parameters were selected in such a way that they gave the best accuracy.

properties. Supporting evidence for the biological relevance of our approach, indicating that our method can extract significant rules only if biologically valid and relevant data is given, can be shown as follows.

First, high prediction performances on diverse datasets might support the validity of our approach. In several datasets consisting of known pair of proteins, including nuclear receptors, GPCRs, ion channels, enzymes and drugs and random protein-drug pairs, our statistical approach with SVM showed high prediction performances (Table III.6). The fact that more than 0.85 AUC and an accuracy of 80% were obtained for diverse datasets suggests that it is possible to extract some properties accountable for interactions between proteins and drugs by statistical approaches.

Second, we showed the biological relevance of these high prediction performances by calculating the prediction performances using biologically meaningless artificial datasets as positives. Several datasets which contained fractions of valid samples found in the DrugBank2 dataset, and which comprised artificial pseudo-positive samples of protein-chemical pairs produced by shuffling with the same frequency of chemical compounds and proteins as that in the DrugBank2 dataset, were generated. Our method was applied to these shuffled artificial datasets (Fig. III.11). Here, if our approach did not depend on the biological properties of the given dataset but only succeeded in classifying given pairs comprising a protein and a chemical compound and random pairs derived from them, the prediction accuracy for each shuffled dataset

Table III.6 Prediction performances in several datasets

dataset	network	# of proteins	# of drugs	# of positives	# of negatives
[1] Nuclear Receptor	Fig. III.12A	26	54	90	200
[1] GPCR	Fig. III.12B	95	223	635	1300
[1] Ion Channel	Fig. III.12C	204	210	1476	3000
[1] Enzyme	Fig. III.12D	664	445	2926	5900
[1] All	Fig. III.12E	989	791	5127	10300
[2] DrugBank2	Fig. III.12F	456	964	1731	3500

dataset	AUC	sensitivity (%)	precision (%)	specificity (%)	accuracy (%)
Nuclear Receptor	0.862	54.4	77.8	93.0	81.0
GPCR	0.906	75.3	81.6	91.7	86.3
Ion Channel	0.943	85.3	85.9	93.1	90.5
Enzyme	0.951	84.2	88.3	94.5	91.1
All	0.956	85.2	88.3	94.4	91.3
DrugBank2	0.882	70.1	82.2	92.5	85.1

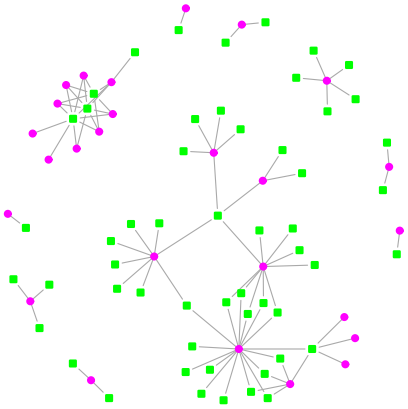
dataset	*AUC	*sensitivity (%)	*precision (%)	*specificity (%)	*accuracy (%)
Nuclear Receptor	0.841	57.8	70.3	89.0	79.3
GPCR	0.893	73.4	77.8	89.8	84.4
Ion Channel	0.931	83.3	82.2	91.1	88.6
Enzyme	0.943	83.1	85.2	92.8	89.6
All	0.948	84.0	85.5	92.9	90.0
DrugBank2	0.877	72.3	78.6	90.3	84.3

[1]: based on Yamanishi *et al.*, 2008.

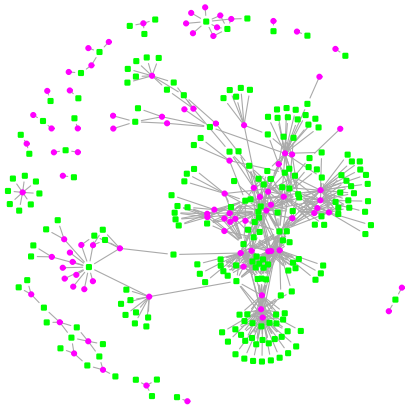
[2]: based on Wishart *et al.*, 2008.

*: RBF kernel was applied to the feature vector concatenating that of protein and that of chemical compound. The results suggest that consideration of combination effects (Eq.(III.12)) generally improves prediction performances.

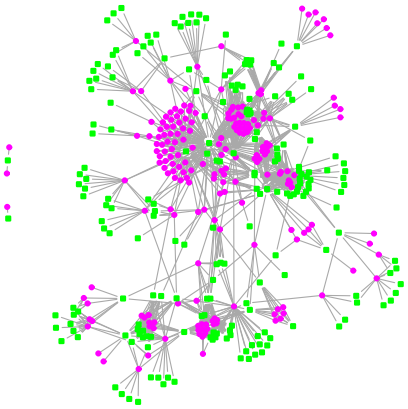
(A) Nuclear Receptor



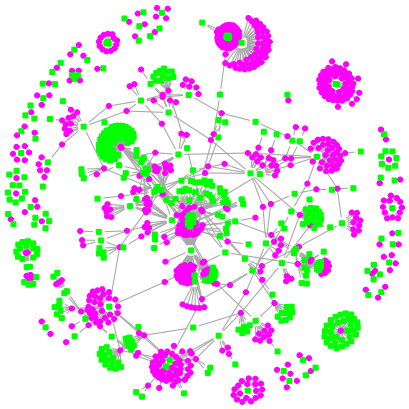
(B) GPCR



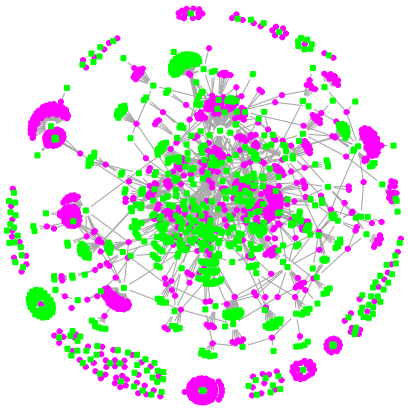
(C) Ion Channel



(D) Enzyme



(E) All



(F) DrugBank2

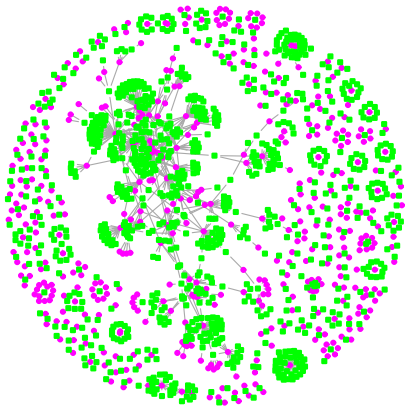


Fig. III.12 Protein-drug interaction network for each dataset. A magenta circle represents a protein and a green square corresponds to a chemical compound. A gray edge means that a drug binds to a protein.

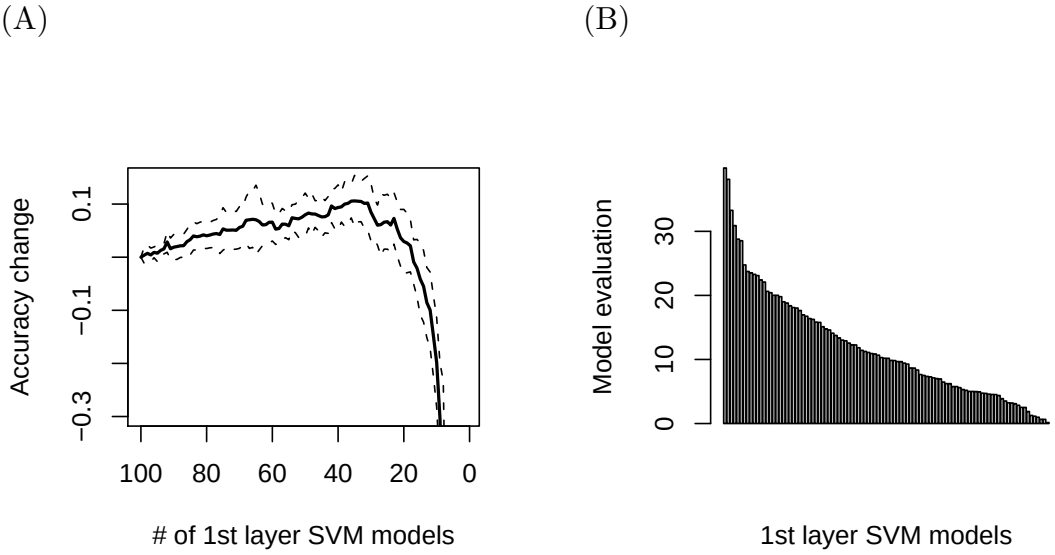


Fig. III.13 Effects of feature selection on two-layer SVM model. (A) Accuracy change according to the number of *subpos* first-layer SVM models used to generate the two-layer SVM model. The accuracy obtained by using 100 first layer SVM models (Sec. 2.3.4) and 10 fold cross validation is set for zero, which corresponds to more than 99% accuracy. The solid line shows the mean of 12 two-layer SVM models. Two dotted lines mean 95% confidence intervals on the assumption of normal distribution. (B) Evaluation of first-layer SVM models based on the Eq. (III.17).

was assumed not to fluctuate.

As shown in Fig. III.11, the prediction accuracy was proportional to the content rate of the biologically valid samples. Therefore, the classification of our approach was shown to function only when a certain amount of biologically valid pairs comprising a protein and a chemical compound are given. This result suggests that our statistical approach succeeds in extracting the rules which are only relevant for the biological binding properties.

3.2 False positive reduction by two-layer SVM and negative data design

3.2.1 Construction of two-layer SVM model

We generated twelve two-layer SVM models (three types of different random pairs as negatives each for four negative sample numbers; 3,500, 7,000, 10,500 and 14,000) to verify the effects of feature selection (Fig. III.13). As shown in Fig. III.13A, reduction of *subpos* first-layer SVM models generally elevates and maintains high prediction accuracy. As far as second-layer SVM models utilized more than 10 first-layer SVM models, changes of prediction accuracy were limited to $\pm 0.2\%$.

These first-layer SVM models were evaluated based on results of feature selection

Table III.7 Prediction performances on different designed negatives

*dataset type	sens.(%)	prec.(%)	acc.(%)	AUC	MCC
min	94.51	96.69	99.42	0.9916	0.9521
mlt	88.68	96.42	99.04	0.9941	0.9189
mle	89.08	96.01	99.04	0.9948	0.9192
max	80.47	88.05	97.99	0.9831	0.8329

*: refers to negative data expansion rules in Sec. 2.4. 10 *subpos* first-layer SVM models were utilized to produce the two-layer SVM model from datasets with 24,500 negatives.

by the following Eq. (Fig. III.13B).

$$\text{model\_evaluation}(i) = \sum_{c=1}^{12}\sum_{j=1}^{100}(acc_{cj}/j) \times s_j(i), \quad s_j(i) = \begin{cases} 1 & \text{if } \ i \in M_{cj} \\ 0 & \text{otherwise,} \end{cases} \tag{III.17}$$

where acc_{cj} is the accuracy obtained by applying j first-layer SVM models to the dataset c, and M_{cj} is a set of these j models. Fig. III.13B shows that contribution to classification markedly differs among the models despite a variety of negative data. This suggests that there exist some combination of first-layer SVM models applicable to a wide range of datasets with no or small loss of prediction performances.

3.2.2 Construction of designed negative data

Based on the preceding findings, we selected 35 first-layer SVM models, which gave the largest accuracy elevation in Fig. III.13A, according to the descending order of the model evaluation in Fig. III.13B. These 35 first-layer SVM models were applied to all the possible combinations of proteins and drugs in the DrugBank2 dataset. The method described in Sec. 2.4 was applied to the yielded $\mathcal{X} = \{\boldsymbol{x}_1 = (p_{1,1}, \ldots, p_{1,35}), \ldots, \boldsymbol{x}_{439584}\}$.

Table III.7 shows prediction performances on datasets produced according to four different rules described in Sec. 2.4. As shown in Table III.7, the dataset *min* gave the best accuracy and the dataset *max* did the worst. As the expanding rule *max* in Sec. 2.4 is designed to construct the dataset which is the most difficult to be classified and the rule *min* is intended for the easiest dataset, these results are reasonable enough.

3.2.3 Evaluation of two-layer SVM and negative data design

Though Table III.7 exhibits very high performances, these results are possibly overestimated as the same positives and negatives were allowed in datasets used to construct first-layer SVM models and second-layer SVM models. Therefore, we tested our prediction method on an external dataset (Table III.8).

In Table III.8, the external dataset consisted of 170 positives and 2,450 negatives that were randomly chosen from 1,731 positives and 24,500 designed negatives with the *mlt* rule and that were excluded in constructing first-layer and second-layer SVM models. Here, "one layer" SVM model was produced based on the same features as used for the first-layer SVM model. To construct the second-layer SVM model, 11

Table III.8 Evaluation of our prediction method on an external test set

	†Internal 10-fold cross-validation			†External prediction		
‡Model type	prec.(%)	sens.(%)	acc.(%)	prec.(%)	sens.(%)	acc.(%)
(A) Effect of rational negative design						
one layer	71.76	42.99	95.11	64.66	50.59	95.00
one layer$_r$	82.38(±0.64)	38.22(±0.95)	95.38(±0.06)	40.68(±1.19)	50.00(±1.87)	92.02(±0.28)
(B) Effect of second-layer SVM model						
subpos	97.11	92.57	99.33	82.81	31.18	95.11
subpos$_r$	94.40(±0.67)	96.46(±1.00)	99.39(±0.11)	42.5(±2.71)	33.53(±3.90)	92.74(±0.41)
subpos$_{\mathrm{v}0.5}$	–	–	–	8.89	57.06	59.27
subpos$^f_{\mathrm{v}0.5}$	–	–	–	8.98(±0.16)	58.90(±1.03)	58.53(±0.39)
subpos$_{\mathrm{v}0.8}$	–	–	–	28.13	5.29	92.98
subpos$_{\mathrm{m}0.5}$	–	–	–	9.52	58.24	61.37
subpos$^f_{\mathrm{m}0.5}$	–	–	–	9.35(±0.14)	59.47(±1.16)	59.94(±0.33)
subpos$_{\mathrm{m}0.8}$	–	–	–	17.86	2.94	92.82
subpos$_{\mathrm{ann}}$	95.98	93.21	99.29	75.81	27.65	94.73
subpos$_{\mathrm{qda}}$	70.69	54.39	95.49	34.52	17.06	92.52
subpos$_f$	95.66(±0.32)	78.33(±1.60)	98.33(±0.10)	78.76(±2.86)	25.59(±1.09)	94.71(±0.09)
(C) Improvement of precision						
allpos	99.68	100.00	99.98	100.00	10.59	94.20
subpos$_{\lfloor 0.9 \rfloor}$	–	–	–	90.70	22.94	94.85
subpos$_{\lfloor 0.95 \rfloor}$	–	–	–	92.50	21.76	94.81
one layer$_{\lfloor 0.9 \rfloor}$	–	–	–	86.67	15.29	94.35
one layer$_{\lfloor 0.95 \rfloor}$	–	–	–	71.43	2.941	93.63

†: The external dataset consisted of 170 positives and 2,450 negatives that were randomly chosen from 1,731 positives and 24,500 designed negatives with the *mlt* rule and that were excluded in constructing first-layer and second-layer SVM models. Here, "one layer" SVM model was produced based on the same features as used for the first-layer SVM model. The internal dataset utilized 1,561 positives and 22,050 negatives.

‡: "Model type" exhibits that the one layer SVM model or the second-layer SVM model, specified by the type of first-layer SVM model, was utilized. Subscripts mean as follows.

- $_r$ means that three types of randomly chosen 22,050 pairs of protein and chemical compound were used instead of designed negatives to construct the SVM model. The 95% confidence intervals are shown.
- $_{\mathrm{v}t}$ means that voting with 11 first-layer SVM models with threshold t was used for prediction.

$$pred_{\mathrm{v}t} = \begin{cases} \text{pos.} & \text{if } \sum_{i=1}^{11} r_i \geq 6 \\ \text{neg.} & \text{otherwise} \end{cases}, \quad r_i = \begin{cases} 1 & \text{if (probability output of model } i) \geq t \\ 0 & \text{otherwise} \end{cases}$$

- $_{\mathrm{m}t}$ means that average of outputs of 11 first-layer SVM models was used for prediction with threshold t.

$$pred_{\mathrm{m}t} = \begin{cases} \text{pos.} & \text{if } \sum_{i=1}^{11} \frac{\text{(probability output of model } i)}{11} \geq t \\ \text{neg.} & \text{otherwise} \end{cases}$$

- $_{\mathrm{ann}}$ means that Artificial Neural Network (ann) (Ripley, 1996) (implemented by the statistical software package R (`http://cran.r-project.org/`) function *nnet*) was applied to outputs of 11 first-layer SVM models. Parameters were selected to give the best accuracy in the internal 10-fold cross validation. For example, 17 units were used in the hidden layer.
- $_{\mathrm{qda}}$ means that Quadratic Discriminant Analysis (qda) (Ripley, 1996) (implemented by R function *qda*) was applied to outputs of 11 first-layer SVM models.
- $_f$ means that twenty types of randomly chosen 11 first-layer SVM models were used to construct the second-layer SVM model. The 95% confidence intervals are shown.
- $_{\lfloor t \rfloor}$(e.g. $t = 0.9$) means that final probability outputs were evaluated with threshold t.

subpos or *allpos* first-layer SVM models, which were chosen by the feature selection method described in Sec. 2.3.4, were utilized.

We evaluate prediction performances with default or higher threshold as prediction results of comprehensive application are often processed with such thresholds. Among several evaluation measures, precision, or reliability of positive prediction, is relevant for measuring the effects of false positive reduction.

As shown in Table III.8A, designed negatives contribute to better prediction performances. Use of rationally designed negatives instead of randomly chosen negatives significantly improved precision of external prediction by more than 20%. Comparison with *subpos* and $subpos_r$ in Table III.8B also exhibited improvement of precision by introducing designed negatives.

Table III.8B shows effectiveness of utilizing the second-layer SVM model. Compared with simplest ways to integrate outputs of several first-layer SVM models including voting ($subpos_{\mathrm{v}}$) and averaging ($subpos_{\mathrm{m}}$), the use of second-layer SVM (*subpos*) model gave highly better overall prediction performances. Besides, though the use of higher threshold leads to better precision at the risk of sensitivity, $subpos_{\mathrm{v}_{0.8}}$ and $subpos_{\mathrm{m}_{0.8}}$ still exhibited lower precision than the second-layer SVM model. Here, the higher threshold (e.g. $subpos_{\mathrm{v}_{0.9}}$) produced no positives.

Other statistical classification methods can be applied to outputs of the first-layer SVM models. In comparison to other major non-linear classification methods including artificial neural networks ($subpos_{\mathrm{ann}}$ in Table III.8B) and quadratic discriminant analysis ($subpos_{\mathrm{qda}}$) (Ripley, 1996), the use of SVM (*subpos*) exhibits better performances in external prediction. These findings show the effectiveness of utilizing SVM to process outputs of the first-layer SVM models.

As described in Sec. 2.3.4, we conducted feature selection in constructing the second-layer SVM model. The observation that the second-layer SVM model with feature selection (*subpos* in Table III.8B) exhibited, in external prediction, significantly higher precision (P-value $= 0.0081$ by t test) and sensitivity (P-value $= 1.8 \times 10^{-9}$ by t test) than that with randomly selected first-layer SVM models supports utilization of the feature selection.

Table III.8C exhibits that the second-layer SVM approach improved precision and suggests that use of *allpos* first-layer SVM models lead to significant reduction of false positives. The second-layer SVM model with *allpos* first-layer SVM models (*allpos* in Table III.8C) achieved 100% precision though prediction sensitivity was low. In general, higher precision can be achieved at the risk of sensitivity by using higher threshold. However, even with higher threshold (e.g. $t = 0.9$), 100% precision weren't realized by the one layer SVM and the second-layer SVM with *subpos* first-layer SVM models (one $\mathrm{layer}_{0.9}$ and $subpos_{0.9}$). Therefore, the use of first-layer SVM models, particularly *allpos* models, in the two-layer SVM approach is considered to contribute to meaningful improvement of precision and reduction of false positives.

Table III.9 Evaluation of our method with respect to comprehensive interaction prediction

[1]dataset	[2]neg.	[3]firsts	[4]P10275	[4]P11229	[4]P35367	$rec_{0.5}$(%)	$rec_{0.95}$(%)	evaluation
(A) one layer SVM								
mlt	16	–	714	1408	1187	100	98.97	82.50
mlt	14	–	709	1820	1634	100	*97.94	*79.02
max	16	–	4073	5956	6964	82.47	*56.70	*47.51
random	14	–	1896.7(±53.6)	10627.3(±648.9)	10204.0(±640.7)	100	99.66(±1.09)	69.20(±0.57)
random	16	–	1869.3(±136.1)	10503.3(±1250.7)	9305.3(±517.8)	100	99.66(±1.09)	69.45(±0.32)
(B) two-layer SVM-*subpos*								
mlt	14	10	177	535	451	96.91	93.81	75.56
mlt	14	11	205	671	491	96.91	91.75	73.54
mlt	14	9	239	513	403	95.88	91.75	73.87
mlt	14	8	290	456	363	88.66	82.47	66.58
mlt	12	10	224	561	612	95.88	92.78	73.25
mlt	16	10	162	466	415	94.85	89.69	73.47
min	14	10	2525	6098	3326	97.94	96.91	69.52
mle	14	10	168	526	599	97.94	92.78	74.79
max	14	10	32	386	191	92.78	*85.57	*72.27
random	14	10	848.3(±345.0)	1531.7(±628.9)	988.0(±411.4)	96.56(±2.89)	81.10(±19.44)	66.44(±7.82)
(C) two-layer SVM-*allpos*								
max	16	9	28	231	129	100	97.94	82.92
max	16	10	29	238	131	100	98.97	82.73
max	16	8	29	243	133	100	96.91	82.09
max	14	9	29	243	129	100	96.91	82.00
mle	16	9	28	267	140	100	100	80.99
mlt	16	9	67	248	141	100	100	80.72
random	16	9	74.7(±42.6)	255.3(±32.2)	146.7(±8.3)	100	100	80.67(±0.93)
(D) only compound SVM[5]								
–	–	–	640	1791	838	86.60	71.13	59.66
(E) similarity search[6]								
–	–	–	1869	1816	1580	–	–	–

[1] refers to negative data expansion rules (details are provided in Supplementary Material). "random" indicates that three types of random pairs comprising a protein and a drug are used as negatives. The 95% confidence intervals are shown.
[2]: the number of negatives (= 1,750×x).
[3]: the number of the first-layer SVM models utilized for the construction of the second-layer SVM model.
[4]: target proteins whose ligands were predicted on the basis of 109,841 compounds. The number of predicted binding compounds is shown.
[5]: SVM model in which chemical compounds binding to each target protein were treated as positives and all other compounds in the DrugBank2 dataset were regarded as negatives.
[6]: A chemical compound i was predicted as a binding ligand of a protein α by using the similarity method if $pred_{\mathrm{sim}}(i) = \max_{j \in A} |I \cap J|/|I \cup J| \geq 0.9$, where A represents the known binding ligands of the protein α, and I (or J) represents a set of substructures considered in calculating the feature vector of the chemical compounds.
*: the threshold was set to 0.9 instead of 0.95 for the calculation of "evaluation" (Eq (III.14)).

3.3 False positive reduction in comprehensive prediction

It is often observed that although statistical learning approaches achieve very high prediction performances in given datasets, statistical prediction models suffer from the problem of generating vast prediction sets including many false positives when applied to a huge dataset such as the PubChem database. In our approach, SVM

models based on feature vectors directly representing amino acid sequences, chemical structures, and random protein-compound pairs as negatives also produced many predictions and inevitably yielded many false positives (Table III.9A *random*).

Upon the introduction of the two-layer SVM and the negatives designed to overcome this drawback, the prediction precision, or the confidence of positive prediction, was significantly improved in computational experiments based on the DrugBank2 dataset (Table III.7 and III.8). The high precision contributes to the selection of more reliable predictions and thus to the reduction of the number of false positives.

Following these results on given datasets, our approaches were evaluated with respect to comprehensive binding ligand prediction. For three proteins (P10275 (androgen receptor), P11299 (muscarinic acetylcholine receptor M1) and P35367 (histamine H1 receptor), their binding ligands were predicted from PubChem Compound 0000001–00125000 which contains 109,841 compounds (Table III.9). Here, P35367 and P11299 are the two most frequently targeted proteins in the DrugBank2 dataset, and P10275 is a protein of average appearance in the DrugBank2 dataset. Among the 109,841 compounds, 47, 45, and 5 known ligands were included for P35367, P11299, and P10275, respectively.

As shown in Table III.9A, utilization of designed negative data (e.g. *mlt* dataset) decreased a number of predicted compounds with a slight loss of sensitivity. Therefore, it is considered to contribute to significant reduction of false positives while detecting the majority of true positives.

In comparison of results for *random* datasets in Table III.9A, III.9B and III.9C , introduction of the two-layer SVM method also shows effects on reduction of false positives. The utilization of *allpos* first-layer SVM models (Table III.9C) decreased a number of predicted compounds more significantly without loss of sensitivity than that of *subpos* first-layer SVM models, which showed a small loss of sensitivity.

Furthermore, as shown in Table III.9A, III.9B and III.9C, the use of carefully selected negatives, the introduction of the two-layer SVM, and the integration of these two approaches efficiently reduced the number of predictions and thus the number of false positives. For example, in comparison to Table III.9A and III.9C, the number of candidates discovered by using the *max* dataset in the *allpos* two-layer SVM approach was about one fiftieth of the number of chemical compounds predicted by using the *random* negative dataset in the one-layer SVM. Furthermore, in comparison to other approaches based on the use of only chemical compounds (Tables III.9D and III.9E), our approaches gave a reasonable number of predictions.

The degree of reduction, sensitivity and precision depend on datasets. For example, use of the dataset *min* gave more predicted compounds than random datasets as it was constructed from positives and negatives that were distant from each other (Table III.9B). On the other hand, utilization of the datasets *mlt*, *mle* or *max* reduced candidate compounds. These datasets were constructed by taking in samples that were difficult to classify by existing SVM models as negatives, and expected to reduce false positives by forming stringent classification boundaries. This result suggests that such concept for negative data design contribute to false positive reduction in comprehensive application of classification methods.

About a number of the first-layer SVM models and negatives, around 10 and around mid-tens-fold of positives respectively were appropriate in this experiment. These values may differ among applications.

According to our proposed measure, one-layer SVM using the *mlt* dataset with 28,000 negatives, two-layer SVM using 10 *subpos* first-layer SVM models and the *mlt* dataset with 24500 negatives and two-layer SVM using 9 *allpos* first-layer SVM models and the *max* dataset with 28,000 negatives can be candidates for the comprehensively applicable protein-chemical interaction prediction model.

Moreover, in comparison to other approaches based only on the properties of chemical compounds (Tables III.9D and III.9E), our approaches gave a reasonable number of predictions.

These results suggest that our prediction models select a reasonable number of ligand candidates from all chemical compounds in large databases, and encourage the comprehensive binding ligand prediction for the target protein.

3.3.1 Comparison with the negative data design on the basis of one-class SVM

One-class SVM (Scholkopf *et al.*, 2001) estimates high-density regions from data samples and is used to detect outliers. A one-class SVM model trained on positive samples can be applied to unspecified samples in order to estimate plausible negatives by selecting samples from outliers. On the other hand, choosing negatives from unspecified samples near the high-density region can contribute to formation of stringent classification boundaries.

In comparison of Table III.10B and III.9B, our method of selecting candidates for negative samples, which involved a phased increase of negative samples on the basis of generated classification boundaries, outperformed that of using the one-class SVM.

3.3.2 Comparison with other prediction approaches

Most previous methods are based on only information of chemical compounds and similarity searches. While our proposed SVM model can deal with any proteins by itself, it is necessary to construct an SVM model for each protein with known binding ligands as positives.

In Table III.9D, "only compound SVM" models for each protein were produced by using D_2^6 in Eq. (III.8) to represent chemical compounds. Here, the same mapping was used for "only compound" and first-layer SVM models to evaluate classification setting itself. It was based on datasets constituted with known approved drugs for each protein as positives and other chemical compounds found in the dataset of binding pairs as negatives. As shown in Table III.9D, these "only compound SVM" models exhibited worse prediction performances than our proposed methods. These findings show that consideration of a pair of a protein and a chemical compound can be useful in predicting binding ligands for proteins.

Moreover, Table III.9E exhibits predictions made by a similarity search. Here, a

Table III.10 Utilization of one-class SVM in the selection of negative samples

(A) One-class SVM models constructed from the DrugBank2 dataset.

[1]kernel	10-fold cross validation accuracy (%)	[2]self-prediction accuracy
sigmoid	87.2	92.1
RBF	50.1	50.0
polynomial	50.0	50.0

(B) Results of comprehensive prediction

[3]kernel	[4]rule	[5]accuracy	[6]P10275	[6]P11229	[6]P35367	[7]evaluation
RBF	*min*	99.9	81317	40897	35466	27.2
RBF	*mle*	99.9	4470	10977	22240	61.7
sigmoid	*min*	99.8	79553	54180	48262	11.9
sigmoid	*mle*	97.8	16194	3316	392	13.5

[1]: Kernel functions used to train one-class SVM models. RBF kernel ($f(\boldsymbol{x}, \boldsymbol{y} = \exp(-\gamma|\boldsymbol{x} - \boldsymbol{y}|^2)$), sigmoid kernel ($f(\boldsymbol{x}, \boldsymbol{y}) = \tanh(\gamma \boldsymbol{x}^t \boldsymbol{y} + c)$), and polynomial kernel ($f(\boldsymbol{x}, \boldsymbol{y}) = (\gamma \boldsymbol{x}^t \boldsymbol{y} + c)^d$). Here, γ, c and d are constants.
[2]: accuracy when a prediction model was constructed from the whole DrugBank2 dataset and applied to the DrugBank2 dataset.
[3]: kernel functions used to train the second-layer SVM model.
[4]: selection rule of candidates for negative data. The one-class SVM model using the sigmoid kernel (A) was applied to all the combinations of proteins and chemical compounds in DrugBank2 dataset except positives, which were represented as described in Sec. 3.2.2, and negative samples were selected according to the following selection rules.

min: Top n samples in the ascending order of the decision values.
mle: Top n samples in the descending order of the decision values whose decision values were under 0.

Here, if the decision value for a sample was above 0, the sample was supposed to be included in the high-density region which was estimated from positive samples. There existed no samples whose decision value equaled to 0.
[5]: 10-fold cross-validation accuracy of the second-layer SVM. 24,500 negatives and 10 *subpos* first-layer SVM models were used to train the second-layer SVM model.
[6]: target proteins whose ligands were predicted on the basis of 109841 compounds. The number of predicted binding compounds is shown.
[7]: refer to Eq.(III.14).

chemical compound i is predicted as a binding ligand of a protein α if

$$pred_{\mathrm{sim}}(i) = \max_{j \in A} \frac{|I \cap J|}{|I \cup J|} \geq 0.9,$$

where A is known binding ligands of the protein α, and I (or J) is a set of substructures in $\mathcal{P}_2^6(i)$ (or $\mathcal{P}_2^6(j)$) described in Eq. (III.8). Compared with Table III.9E, our proposed methods (e.g. the two-layer SVM in Table III.9B, C) gave a smaller number of candidates, which included chemical compounds that weren't found by the similarity search (Table III.11). These results suggest that, in terms of selection of, preferably novel, candidates to be experimentally verified and reduction of false positives, our proposed methods can achieve these purposes more efficiently than mere

Table III.11 Overlaps of predictions between prediction models in Table III.9

(A) threshold= 0.5

	1.	2.	3.	4.	5.
1. [a]similarity	5265	1161	1145	654	354
2. only compound	–	3269	2007	658	345
3. [b]one layer	–	–	3309	[†] 895	388
4. [c]*subpos*	–	–	–	1163	309
5. [d]*allpos*	–	–	–	–	388

(B) threshold= 0.95

	1.	2.	3.	4.	5.
1. similarity	5265	359	260	337	253
2. only compound	–	456	171	200	176
3. one layer	–	–	268	175	200
4. *subpos*	–	–	–	428	179
5. *allpos*	–	–	–	–	265

[†]: There were 895 compounds common between 3,309 compounds predicted by the one layer SVM model and 1,163 obtained by the *subpos* two-layer SVM model.
[a]: A chemical compound i is predicted as a binding ligand of a protein α by the similarity method if

$$pred_{\mathrm{sim}}(i) = \max_{j \in A} \frac{|I \cap J|}{|I \cup J|} \geq 0.9,$$

where A is known binding ligands of the protein α, and I (or J) is a set of substructures in $\mathcal{P}_2^6(i)$ (or $\mathcal{P}_2^6(j)$) described in Eq. (III.8).
[b]: one layer SVM using the *mlt* dataset with 28,000 negatives.
[c]: two-layer SVM using 10 *subpos* first-layer SVM models and the *mlt* dataset with 24500 negatives .
[d]: two-layer SVM using 9 *allpos* first-layer SVM models and the *max* dataset with 28,000 negatives.

similarity searches.

3.3.3 Overlaps of predictions between prediction models

Table III.11 exhibits overlaps of predicted chemical compounds between prediction models. Our proposed methods shared more than half of their predictions with the only compound SVM. Therefore, consideration of pairs is relevant to utilization of only chemical compound information, or well used ligand based virtual screening approach. With differences of candidates that they cover, they can complement each other.

Our methods aimed to reduce false positives, and thus, as shown in Table III.11, predicted a smaller number of candidates than mere similarity search, which is based on the assumption that similar compounds have similar functions. The observation that more than half of candidates predicted by our method were structurally similar to known binding ligands and that the ratio increased with higher threshold partly shows validity of our prediction models, and effects of our methods on efficient candidate selection. On the other hand, the fact that even the least productive or the most stringent model (*allpos*) gave candidates that weren't similar to known ligands implies

Table III.12 Genome-wide target protein prediction

A. Predicted target proteins of MDMA

[1]ID	description	[2]prob.
P08588	Beta-1 adrenergic receptor	0.956
P23975	Sodium-dependent noradrenaline transporter	0.930
P31645	Sodium-dependent serotonin transporter	0.930
P03372	Estrogen receptor	0.905
P35348	Alpha-1A adrenergic receptor	0.905

B. Prediction of interaction between MDMA and adrenergic receptor α1 subfamily members

protein	prob.	[3]agonist prob.
ADRα1A	0.722	0.982
ADRα1B	0.778	0.982
ADRα1D	0.802	0.982

[1]: UniProt ID. [2]: Estimated binding probability by an SVM model. [3]: the estimated probability that a compound acts as an agonist calculated by a SVM model.

possible contribution of our methods to discovery of completely novel ligands.

3.4 Utilization of feedback and additional data

The experimental verification of the computational predictions produces feedback data or samples which are not included in the given training datasets. The efficient utilization of these data can contribute to the fast identification of compounds with the desired properties and can be an advantage of statistical learning approaches.

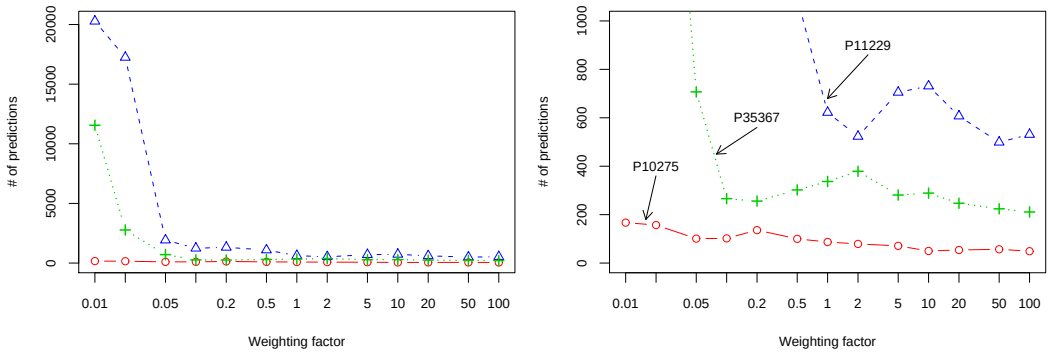
As shown in Fig. III.14A and B, the use of the additional model with a sufficient weighting factor controlled the increase of the predictions with a slight decrease of the recall rate. Although using large weighting factors relatively decrease the influence of other first layer SVM models derived from the DrugBank2 dataset in classification, the low performances of "only additional model:st2" in Fig. III.14, where only one first-layer SVM model derived from additional data was used to construct the second-layer SVM model, indicate necessity of the first-layer SVM models derived from the DrugBank2 dataset and combination of these first-layer SVM models and an additional first-layer SVM model.

With this efficient strategy for utilizing feedback data, computational prediction and experimental verification improve each other to enable faster progress toward the identification of useful small molecules.

(A) Comparison of strategy for utilization of feedback and additional data

[1]strategy	[2]P10275	[2]P11229	[2]P35367	$rec_{0.5}$(%)	$rec_{0.95}$(%)	evaluation
[3]one-layer:st1	1189	6293	2549	100	*100	*74.07
[4]two-layer:st1	174	22160	12821	98.97	97.94	63.08
[4]two-layer:st2	57	499	224	100	98.97	79.22
[5]only compound:st1	521	3985	2563	83.51	23.71	43.57
[6]only involved pairs:st1	503	3600	2390	81.44	25.77	43.04
[7]only additional model:st2	82	6547	4606	30.93	27.84	16.22

(B) The weighting factor controlled the number of predictions
(1) The number of predictions at the threshold of 0.5



(2) The number of predictions at the threshold of 0.95

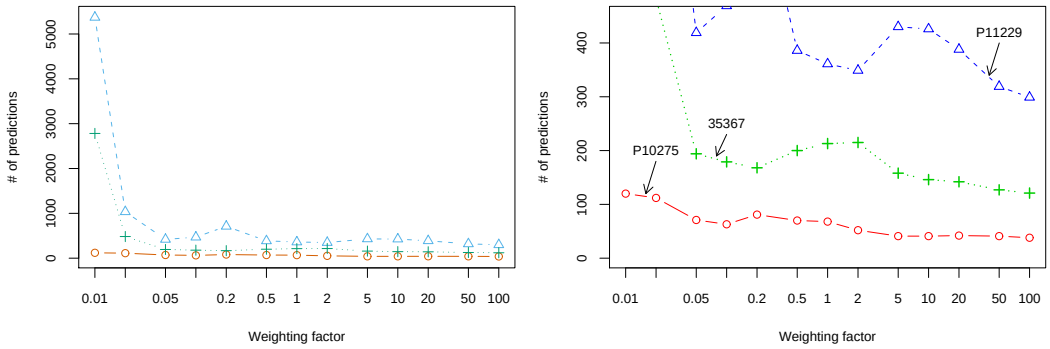


Fig. III.14 Effects of the strategy for the utilization of feedback and additional data. (A) [1]: st1; a strategy where additional data, or pairs comprising a chemical compound and a protein, were simply added to the training samples in constructing a prediction model. st2; a strategy where additional data were first used for the construction of an additional first-layer SVM model and second added to the training samples in the construction of a second-layer SVM model. [2]: target proteins whose ligands were predicted from 109,841 compounds. The number of predicted ligands is shown. [3]: one-layer SVM using the *mlt* dataset with 28,000 negatives. [4]: two-layer SVM using 9 *allpos* first-layer SVM models and the *max* dataset with 28,000 negatives. In st2, the weighting factor was set to 50. [5]: SVM model where the chemical compounds binding to each target protein were treated as positives, and all other compounds in the DrugBank2 dataset were regarded as negatives. [6]: SVM model where pairs of all target proteins and known ligands were treated as positives, while pairs of all target proteins with other compounds were regarded as negatives. [7]: two-layer SVM model in which only one first-layer SVM model derived from additional data was used for the construction of a second-layer SVM model. *: a threshold of 0.9 was used instead of 0.95 for the calculation of "evaluation" (Eq. (III.14)). (B) The relation between the weighting factors and the number of predictions is shown for the case where the threshold= 0.5.

3.5 Genome-wide target protein prediction

3.5.1 Target protein prediction of MDMA

One of the advantages of our proposed method is that screening target proteins for a chemical compound can be performed on a genome-wide scale. This is due to the fact that our method can be applied to all proteins whose amino acid sequences have been determined even though the 3D structural data is not yet available. Furthermore, our method can also be applied to chemical compounds that have been identified by high-throughput analysis using MS (Mass Spectrometry), but whose chemical structures has yet to be determined. These advantages of our novel prediction methodology may therefore facilitate the identification of unknown functions of novel chemical compounds by using their predicted target proteins as characterization profiles. Additionally, further predictions of possible adverse effects of chemical agents may be made by identifying unexpected protein targets.

We conducted genome wide target protein predictions for MDMA from a pool of 13,487 human proteins derived from the UniProtKB/Swiss-Prot protein knowledgebase release 49.0 (Apweiler *et al.*, 2004) (Table III.12A, B, and see Supplementary Table IV). For this purpose, we used our general binding prediction model, exploiting mapping C, F and G with 2,000 artificially generated negative samples. The number of negative samples was set at 2,000 as this gave the best MCC score (Table III.8). MDMA, or ecstasy, is one of the best known psychoactive drugs, but is also believed to be effective in the treatment of post-traumatic stress disorder (PTSD).

MDMA was predicted to bind to 56 different proteins among the 13,487 proteins screened using our model, and the 5 proteins with the highest binding probabilities are listed in Table III.12A. MDMA was correctly predicted to bind to sodium-dependent serotonin transporter (5HTT), and this binding prediction is validated by the existing evidence that MDMA stimulates serotonin secretion and exhibits psychoactivity by binding to 5HTT (Rudnick and Wall, 1992). Moreover, our specific binding prediction model, constructed from the ADR drug dataset, predicted that MDMA binds to the α-1 adrenergic receptor families and activates them (Table III.12B) This is also biologically correct, as MDMA-induced hyperthemia is known to be caused by the activation of α-1 adrenergic receptors, in conjunction with the β-3 adrenergic receptor (Sprague *et al.*, 2003).

It is noteworthy that the known binding of MDMA to β-3 adrenergic receptor is not predicted by our method but this may be due to the lack of positive samples containing this receptor.

Overall, we conclude that our current prediction results indicate the biological plausibility of undertaking genome-wide analyses using our proposed novel method.

3.5.2 Comparisons with the similarity-based search method

Sequence similarities between these predicted target proteins of MDMA were relatively low (Table III.13). For example, 5HTT (P31645) and ADR α-1A (P35348), showed only about 10% sequence similarity though both were reported to interact

Table III.13 Sequence similarities between predicted target proteins of MDMA.Sequence similarities were calculated by ClustalW.

protein	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1 (P08588)	100									
S2 (P23975)	11.1	100								
S3 (P31645)	11.1	45.3	100							
S4 (P03372)	11.3	9.7	9.7	100						
S5 (P35348)	21.0	10.9	10.3	10.5	100					
S6 (P35372)	16.3	10.8	11.5	11.5	18.8	100				
S7 (P41145)	18.4	11.1	12.4	10	18.9	55.5	100			
S8 (P28223)	22.7	11.3	11.5	11.3	19.1	16.3	17.1	100		
S9 (Q01959)	11.3	65.2	44.2	10.1	11.2	12.3	11.6	10.6	100	
S10(P35367)	18.9	10.9	11.1	11.3	18.9	17.8	19.2	16.1	10.1	100

Table III.14 Chemical compounds identified by the similar structure search of MDMA

Approved drug	Similarity score	target proteins
Phenmetrazine	19	NET(P23975), DAT(Q01959), MAOA(P21397)
Pilocarpine	19	MARM1(P11229)
Phensuximide	17	-
Chloroxine	17	-
Ciclopirox	17	ALOX15 (P163050), COX1(P23219),
Tranylcypromine	16	MAOA(P21397)
Ethotonin	16	HH1(Q14524)
Tolazoline	15	AR1A(P35348)
Nicotine	15	CHRNA2(Q15822)
Chlorzoxazone	15	KCNMA1(Q12791)
Ketamine	15	NMDAR-L(Q8TCU5)
Metaxalone	15	-
Primidone	15	GABARA1(P14867)
Methylphendiate	15	DAT(Q01959), NET(P23975)

with MDMA (Rudnick and Wall, 1992; Sprague *et al.*, 2003). On the other hand, the similar chemical structure search of MDMA, which was conducted by the DrugBank web service (Wishart *et al.*, 2008), showed no approved drugs that had 5HTT (P31645) as their target (Table III.14).

These results suggest that our method can identify novel target proteins or chemical compounds that are not similar to known targets and that are not found by similarity-based search methods such as BLAST.

In the researches of protein family detection, it has been shown that the kernel methods such as SVM can detect remote protein evolutionary and structural relationships more sensitively and more specifically than the simple sequence similarity-based method such as PSI-BLAST (Liao and Noble, 2003; Leslie *et al.*, 2004). Therefore, we conclude that the use of the kernel method and the consideration of multiple types of interactions between proteins and chemical compounds are effective in the comprehensive protein-compound interaction prediction.

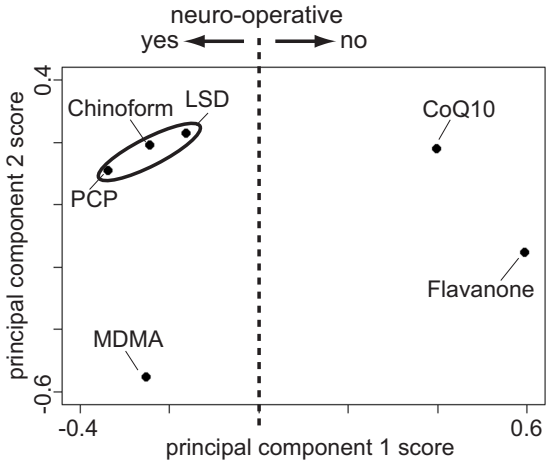


Fig. III.15 The similarity between chemical compounds based upon their target protein profiles. The distance between two compounds is defined based on their predicted target protein profiles (See Methods for details), and principal component analysis (PCA) was then applied to the distance matrices for 6 chemical compounds. Each plot reflects the principal component 1 score and principal component 2 score for each small molecule. The cumulative proportion of these two components is 75.8%.

3.5.3 Interactomical profile

By utilizing genome-wide target protein predictions, it will also be possible to classify chemical compounds according to their predicted protein targets and this profile may also be used to classify their functions.

In this context, we applied principal component analysis (PCA) to the distances between compounds in terms of the overlaps between their target proteins (Fig. III.15). Based upon these PCA results shown in Fig. III.15, it is clear that there are boundaries separating one group of chemical compounds including psychoactive drugs such as LSD, MDMA and PCP and the other groups including coenzyme Q10 and flavanone that have a number of effects in the body but don't act on neural systems. In addition, strong similarities between LSD, PCP and chinoform, which has been reported to cause a serious neuropathy called SMON, are suggested by these analyses.

3.6 Comprehensive binding ligand prediction

3.6.1 Application of our strategy to the discovery of androgen receptor binding ligands

■First computational prediction

We set the human androgen receptor (AR) as the target protein. The two-layer SVM model with an additional model for the androgen receptor was applied to the screening for human androgen receptor binding ligands from 19,171,127 chemical compounds in the PubChem Compound database. As a result, 500 chemical compounds (compounds of the same connectivity were counted only once) were predicted (Fig. III.16B).

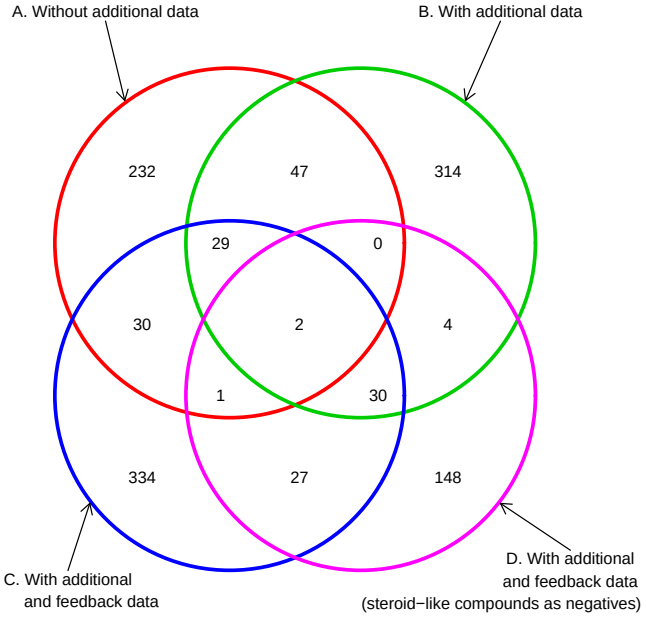
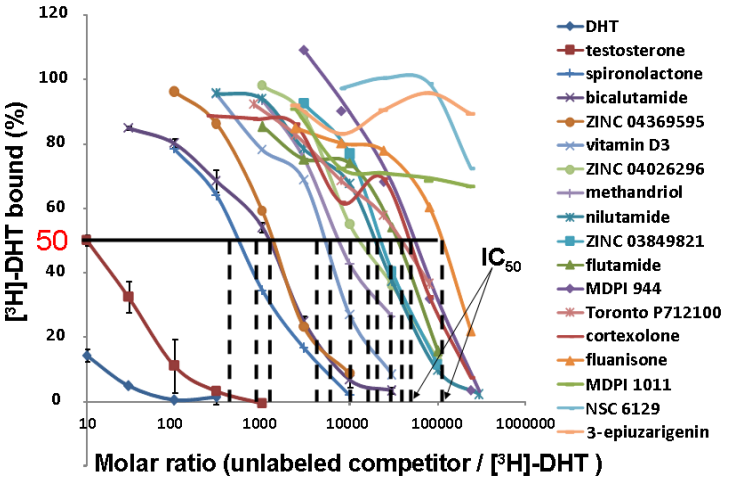


Fig. III.16 The scope of the predictions changed depending on whether additional or feedback data was used. (A) 342 predictions (Supplementary Table A42) made with a two-layer SVM model (= *model A*), which utilized 9 *allpos* first-layer SVM models and 28,000 negatives selected in accordance with the *max* rule. (B) 500 predictions (Supplementary Table A43) made with a model (=*model B*) which utilized one more first-layer SVM model in addition to *model A* derived from chemical compounds known to bind to AR and which were not included in the DrugBank2 dataset. These additional data, or pairs between these compounds and AR, were used as positives in the construction of the second-layer SVM model. (C) 527 predictions (Supplementary Table A44) made with a model (= *model C*) in which feedback from our experiments was used for the production of additional positives and negatives along with the additional data used in constructing *model B*. (D) 213 predictions (Supplementary Table A45) made with a model (= *model D*) in which pairs of chemical compounds with steroid structure and AR were treated as negatives in constructing the additional model and the second-layer SVM model. The weighting factor w was set to 10 and was used for the additional models in *models* B, C and D. Not all overlaps are shown.

■First experimental verification

Out of 500 computationally predicted candidates, an *in vitro* binding assay was applied to 18 purchasable chemical compounds (details are provided in Supplementary Fig. 1), including 6 known drugs or androgens (Fig. III.17). The results obtained for these 6 known ligands agreed well with the results found in the relevant literature (Roselli, 1998), thus proving the reliability of the assay.

For 12 predictions except 6 known ligands by applying a threshold level of IC_{50} = 100 μM, which was based on the fact that IC_{50} of flutamide, a known drug, was more



compound	[1]ID	[2]pred.	[3]IC_{50}
[4]DHT	15	Pos.(97)	0.00374
[4]testosterone	6013	Pos.(97)	0.02
[4]spironolactone	5833	Pos.(97)	1.26
[4]bicalutamide	56069	Pos.(97)	2.54
ZINC 04369595	11890335	Pos.(58)	4.14
vitamin D3	6221	Neg.(< 1)	12.26
methandriol	10632	Neg.(37)	16.92
ZINC 04026296	11872740	Pos.(99)	30.6
[4]nilutamide	4493	Pos.(97)	43.2
ZINC 03849821	11868792	Pos.(80)	49.4
[4]flutamide	3397	Pos.(97)	53.4
MDPI 944	11867734	Pos.(97)	92.4
Toronto P712100	4086221	Pos.(99)	108.2
cortexolone	9050	Neg.(< 1)	131.8
fluanisone	15139	Neg.(< 1)	304
MDPI 1011	2748171	Pos.(68)	> 600
NSC 6129	15503	Neg.(< 1)	> 600
3-epiuzarigenin	10071	Neg.(3.9)	> 600

Fig. III.17 Results of an *in vitro* binding assay. [1]; PubChem Compound ID. [2]; computational prediction, label (predicted probability (%) for a positive outcome). [3]; The concentration (μM) of an unlabeled test compound, in which, according to the measured radioactivity, 50% of the [^{3}H]-DHT is still bound to MBP-ARC. [4]; chemical compounds included in the DrugBank2 dataset or additional data.

than 50 μM, a precision of 67% (4/6) and an accuracy of 67% (8/12) were obtained (Fig. III.17B). Two misclassified compounds which were not detected in our method but proved to bind to the androgen receptor can now be utilized in order to refine the predictions.

■Second computational prediction with feedback By utilizing the results of the first experimental verification, the prediction model was reconstructed. Although the first computational prediction and experimental verification involved many compounds with steroid skeletons, binding of steroid-like compounds to the androgen receptor, which is a steroid-hormone receptor, is relatively obvious. Moreover, since steroid-like compounds are expected to act as agonists of the androgen receptor, antagonists are given preference in terms of search for chemical compounds of potential therapeutic effects for human prostate cancer, which involves activation of the androgen receptor. Thus, the prediction model in which pairs of the androgen receptor and steroid-like chemical compounds were regarded as negatives was also constructed in order to search for antagonists of the androgen receptor. The prediction coverage of these two models (Fig. III.16C and D) was different. The latter prediction models predicted chemical compounds without steroid skeletons, as expected.

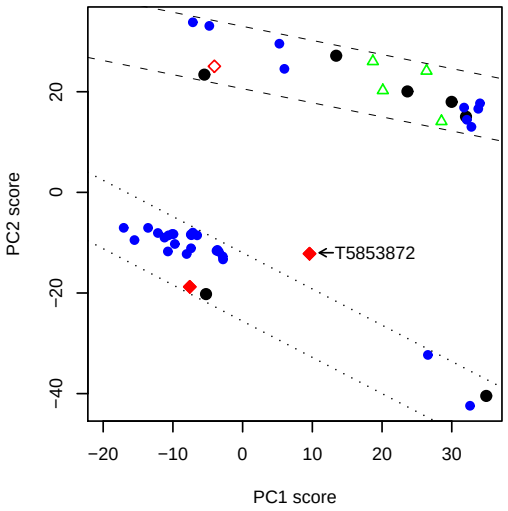
■Second experimental verification

5 candidates predicted with the models reconstructed with feedback data and different strategies as described in the previous section were experimentally verified (details are provided in Supplmentary Fig. 1). Out of these 5 candidates, 3 chemical compounds bound to the androgen receptor at a threshold of 100 μM (Fig. III.18A), thus achieving 60% precision (3/5).

(A) Results of the second experimental verification

compound	[1]ID	[2]pred.	[3]IC_{50}
[c]DSHS00507SC	2807124	Pos.(79)	22
[d]4J-584S	1476447	Pos.(54)	22
[d]T5853872	17566945	Pos.(84)	80
[c]BAS01279920	3129307	Pos.(79)	> 200
[d]AN-562/43163258	838171	Pos.(53)	> 200

(B) Predicted compounds in the chemical space



(C) Chemical structure of T5853872

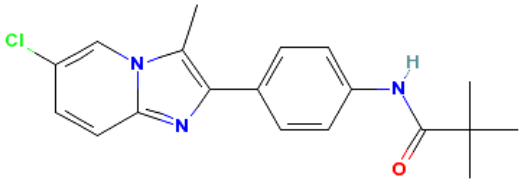


Fig. III.18 The second experimental verification showed about 60% accuracy and the chemical space of verified compounds was explored. (A) [1]; PubChem Compound ID. [2]; computational prediction, label (predicted probability (%) for a positive outcome). [3]; The concentration (μM) of an unlabeled test compound, in which, according to the measured radioactivity, 50% of the [^{3}H]-DHT is still bound to MBP-ARC. [c]; predictions belonging to Fig. III.16C. [d]; predictions belonging to Fig. III.16D. (B) The chemical space based on E-Dragon (Tetko *et al.*, 2005) descriptors and the principal component analysis (PCA) applied to known ligands and additional data. The cumulative proportion of the two components used to draw the plot is 83.4%. The black circles correspond to known ligands in the DrugBank2 dataset, the blue circles represent additional data, the green triangles correspond to true positives in the first computational prediction, the red diamonds correspond to true positives in the second computational prediction. The open red diamond belong to Fig. III.16C, and the solid red diamonds belong to Fig. III.16D. Chemical compounds located between the two dashed lines have steroid-like structures. (C) A potential ligand with a chemical structure differing from the structures of known ligands.

As shown in Fig. III.18B, known drugs and chemical compounds in the additional data can be roughly divided into two regions in the chemical space, which is based on the results of PCA applied to known ligands and chemical compounds in additional data represented by E-Dragon (Tetko *et al.*, 2005) descriptors. Although all true positives of the first computational prediction belonged to one of these regions, T5853872 (* in Fig. III.18B and Fig. III.18C), which is one of the second computational predictions, was not included in these regions. This result suggests that repeating the processes of computational prediction, experimental verification and the feedback of experimental results for new predictions contributes to the efficient exploration of the chemical space targeted in the search as well as to the discovery of novel ligands.

■Third computational prediction with feedback

The results of the second experimental verifications were feedbacked to re-construct

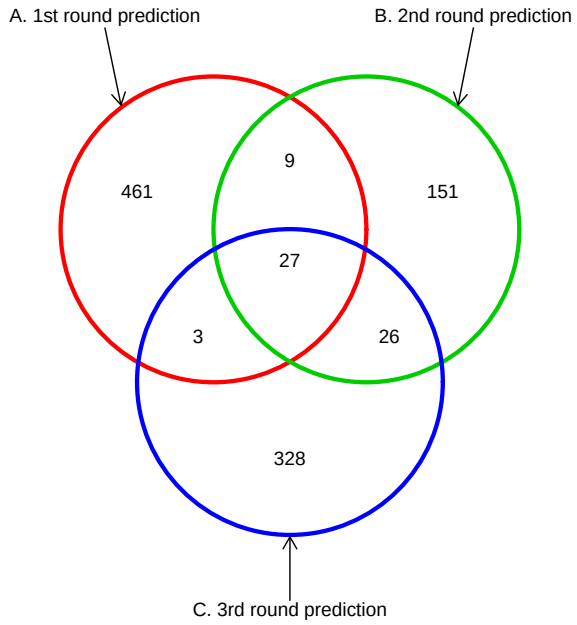


Fig. III.19 The scope of the third computational prediction. A. 500 predictions made by *model B* in Fig. III.16B. B. 213 predictions made by *model D* in Fig. III.16D. C. 384 (Supplementary Table A44) predictions made by the model (= *model E*) in which the results of our second experiment were utilized and pairs between chemical compounds with structure similar to steroid and AR were treated as negatives in constructing the additional model and the second-layer SVM model. There were some compounds like vitamin D3 (Supplementary Fig. 1-VI) that were treated as positives in *model D* but as negative in *model E*. The weighting factor $w = 10$ was used for all the additional models.

the third prediction model. In constructing this prediction model, some compounds that had structures similar to steroid skeletons were also regarded as negatives. As shown in Fig. III.19, this model produced predictions that were not included in the first or second computational predictions. As expected from the design of the prediction model, more compounds different from known drugs were predicted than the first or second computational prediction. Some predicted compounds were more similar to T5853872 (Fig. III.18C) than known drugs or compounds in the additional data. This finding suggests influence of the feedback of our second experiment.

3.6.2 Comparison with the docking analysis

The docking analysis with AutoDock (Morris *et al.*, 1998) was applied to the human androgen receptor ligand-binding domain (PDB code; 2AM9 (Pereira de Jesus-Tran

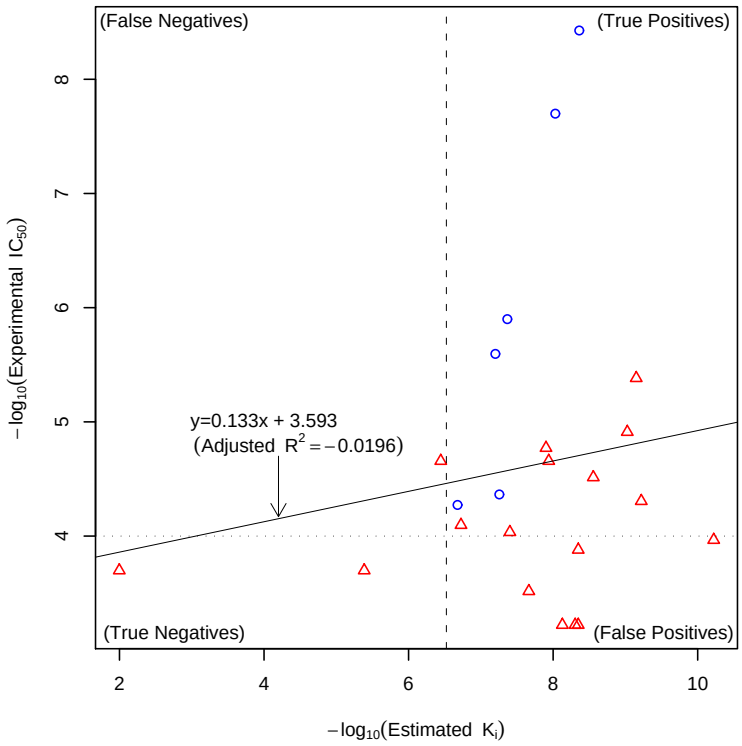


Fig. III.20 The docking analyses of experimentally verified compounds. The docking analysis with AutoDock (Morris *et al.*, 1998) was applied to the human androgen receptor ligand-binding domain (PDB code; 2AM9 (Pereira de Jesus-Tran *et al.*, 2006)) and tested compounds whose 3D structures were generated by obgen in the Open Babel package ver.2.2.0 (Guha *et al.*, 2006) or CORINA (Sadowski and Gasteiger, 1993). The conditions of AutoDock followed to Jenwitheesuk and Samudrala, 2005. ARG752 of 2AM9, which was estimated to be important for the binding of androgens by the human androgen receptor (Pereira de Jesus-Tran *et al.*, 2006), was set to a flexible residue in AutoDock. Blue circles show known compounds. Red triangles are other tested compounds. $-\log$(Estimated K_i) was derived from the estimated inhibition constant of the first cluster in the AutoDock output. The horizontal dotted line is the threshold of 100 μM. The vertical dashed line is the threshold of 300 nM, which is based on the estimated K_i 210.27 nM of flutamide, a known drug.

et al., 2006)) and tested compounds whose 3D structures were generated by obgen in the Open Babel package ver.2.2.0 (Guha *et al.*, 2006) or CORINA (Sadowski and Gasteiger, 1993).

As shown Fig. III.20, for 17 chemical compounds that were experimentally verified in this study, 59% accuracy (10/17) and 57% precision (8/14) were achieved. In this experiment, correlation between the experimental IC_{50} and the estimated K_i was not observed (adjusted R^2 was around zero in Fig. III.20).

As the docikng analysis is a well used method in the virtual screening, this finding that our method (64% precision (7/11) for 17 tested compounds) outperformed AutoDock, one of the most frequently used docikng tools, indicates effectiveness of our method in virtual screening.

4 Discussion

We proposed a comprehensively applicable computational method for predicting the interactions between proteins and chemical compounds, in which the number of false positives was reduced in comparison to other methods. Furthermore, we proposed the strategy for the efficient utilization of experimental feedback and the integration of computational prediction and experimental verification.

The application of our method to the androgen receptor resulted in 67% (4/6) prediction precision according to *in vitro* experimental verification in the first computational prediction and 60% (3/5) in the second prediction, which included the feedback of the first experimental verification. However, these relatively low precision values do not represent the true statistical significance.

This $60 - 70\%$ precision can also be evaluated by using the following P-value.

$$P\text{-value} = \sum_{x=p}^{t} \frac{{}_{M}C_{x} \times_{(N-M)} C_{(t-x)}}{{}_{N}C_{t}}$$

Here, N is the number of prediction targets, M the number of ligands potentially binding to the target proteins, t is the number of tested compounds, and p is the number of true positives. With $N = 19171127$, which is the the number of chemical compounds in the PubChem Compound database and $M = 19171127 \times (456/3000) \times (7/964) \fallingdotseq 21160$, which is based on the optimistic assumption that all compounds can be regarded as potential drugs for some target protein, the estimation that 3000 druggable proteins exist (Russ and Lampel, 2005) and the distribution of target proteins and drugs in the DrugBank2 dataset consisting of 456 target proteins and 964 drugs including 7 known ligands for the human androgen receptor, P-values of 2.21×10^{-11} and 1.34×10^{-8} are obtained for the prediction precision of the first and the second computational prediction, respectively. These extremely small P-values prove the significance of the virtual screening and its precision in the drug discovery process.

These prediction performances are as good as or better than several previous virtual screening studies based mainly on docikng analyses (Cosconati *et al.*, 2008; Finn *et al.*, 2008; Zhong *et al.*, 2008). For example, at a threshold of 100 μM, 7% precision (3/39) for *Mycobacterium tuberculosis* adenosine 5′-phosphosulfate reductase (Cosconati *et al.*, 2008), 71% precision (22/31) for *Staphylococcus aureus* methyonyl-tRNA synthetase (Finn *et al.*, 2008) and 8% precision (16/192) for human DNA ligase VI (Zhong *et al.*, 2008) were obtained, respectively. In addition, for the 17 chemical compounds which were experimentally verified in this study, 59% accuracy (10/17) and 57% precision (8/14) was achieved in the docikng analysis using AutoDock (Morris *et al.*, 1998) (Sec. 3.6.2). Note that the docikng analysis with AutoDock was not applied to the screening for human androgen receptor binding ligands from 19,171,127 compounds in the PubChem Compound database, but applied only to 17 compounds which were the results of virtual screening by our method. In terms of computational time, for binding prediction of one pair of a protein and a

chemical compound, by using one Opteron 275 2.2 GHz CPU, AutoDock took approximately 100 minutes in average with 100 genetic algorithm (GA) runs depending on properties of chemical compounds e.g. the number of rotatable bonds and the number of atom types while our method required less than 0.3 seconds. These computational time comparisons indicate that our method can perform a virtual screening of more than 19 million chemical compounds from the PubChem Compound database for any proteins in genome-wide scale and this immense screening task would be infeasible to accomplish with any of the existing docking methods.

Furthermore, the fact that the second computational prediction, or the use of feedback data, contributed to the discovery of novel ligands (Fig. III.18B) supports the utilization of statistical learning methods in virtual screening.

Regarding the computational prediction method used in this paper, we made the method available to the public as a web-based service named COPICAT (COmprehensive Predictor of Interactions between Chemical compounds And Target proteins; `http://copicat.dna.bio.keio.ac.jp/`).

Although we didn't utilize mass spectra in comprehensive binding ligand prediction due to the lack of huge databases, one-layer SVM with the mass spectrometry data showed the same level of prediction performances with that using the chemical structure data (Table III.7A and III.8). Mass spectrometry data have been rapidly produced by comprehensive metabolite analyses mainly to quantitate known chemical compounds. These analyses have also produced many spectra whose corresponding chemical structure is unknown. Our method could be used to predict functions of these unknown chemical compounds with the profiles of predicted target proteins (Table III.15). In addition, predicted functions would be of use to decide the priority order of determining the chemical structure of unknown spectra. Determined chemical structures, combined with mass spectra, would improve the prediction accuracy (Table III.7A and III.8), and further elucidate the biological roles of chemical compounds.

Moreover, in addition to comprehensive metabolite analyses, MS methods have now been exploited to obtain high-throughput profiles of glycans from cells and tissues (An *et al.*, 2003). This indicated a possible application of a method that incorporated such MS data to the prediction of glycosylation, or the attachment of glycans or carbohydrates to proteins. Since glycosylation is the most significant and active post-translational modification in the cell, this approach could be developed into more precise protein-chemical interaction prediction method to identify unknown functions of small molecules.

In our present study, we used EI mass spectrometry data due to data availability although EI-MS spectra have some weakness of abundance and reproducibility. However, our method is general enough to be applied to MS/MS spectra which show many fragments representing chemical substructures as EI-MS spectra do and which will be produced and accumulated rapidly in the comprehensive metabolite analyses such as CE-MS. Therefore, our approach could be one of effective ways to directly exploit mass spectrometry data that will be produced at ever increasing speed.

Part IV

Discussion

Fig. IV.1 and IV.2 illustrates this study and related studies. In Part II of this thesis, we suggested that there exist relationships between a transcription factor network consisting of transcription factors as a node and cooperativity between them as a edge and a protein-protein interaction network composed of proteins as a node and physical interactions between them as a edge (Fig. IV.1). In particular, the hypothesis that proteins regulated by TFs functioning cooperativity are close to each other in the protein-protein interaction network successfully enabled us to predict cooperative TFs (Table II.1-II.3). These findings can be interpreted that efficiency of transcription is realized by expressing at the same timing proteins that function in the same biological process.

With respect to regulation of protein expression, other possible mechanism than the transcription factor network is regulation by non-coding RNA (ncRNA). Recently, it is discovered that transcripts from mammalian genomes include many ncRNAs (Carninci *et al.*, 2005). These findings suggest that ncRNA plays a important role in regulating the timing and rate of protein translation. For example, micro-RNA (miRNA) binds to mRNA that contains sequences complementary to a part of miRNA and inhibits translation (Lee *et al.*, 1993; Baulcombe, 2002). In general, the higher eukaryotic complexity is, the bigger the fraction of non-coding DNA (Taft *et al.*, 2007). Thus, it is suggested that the higher eukaryotic complexity is, the more important regulation by ncRNA is.

Our experiment using ChIP data and protein-protein interaction data of *S. cerevisiae* didn't consider ncRNAs whose influence may be small in lower eukaryotes. Although relations between transcription factor networks and ncRNAs are not clear, relationships between protein-protein interaction networks and ncRNAs are to be analyzed in order to infer to what extent the timing of protein expression corresponds with the timing when proteins function in biological processes.

In Part III of this thesis, we developed a comprehensively applicable prediction

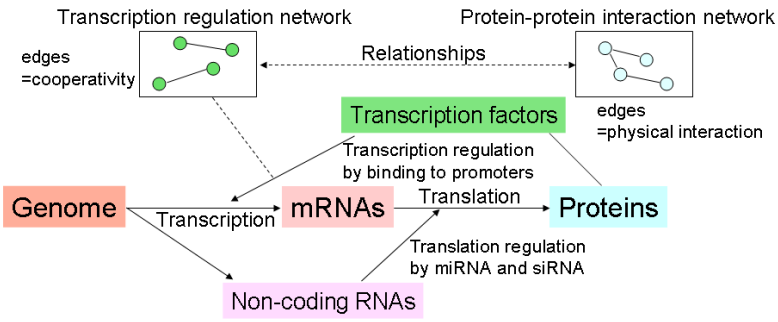


Fig. IV.1 Schematic illustration of mechanisms for regulation of protein expression

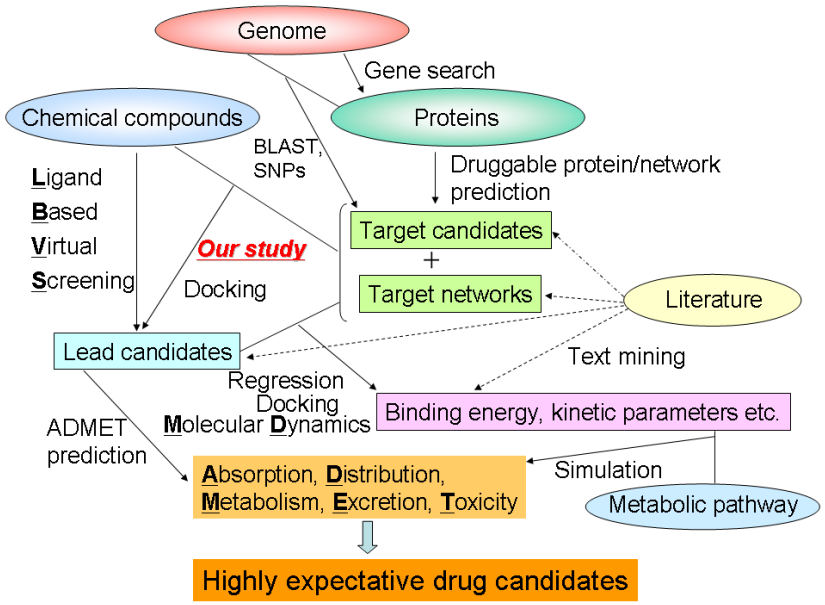


Fig. IV.2 Schematic illustration of possible steps toward and possible contributions of bioinformatics to a discovery of highly expectative drug candidates

method for interactions between proteins and chemical compounds. This method is one of the first steps toward a discovery of highly confident drug candidates (Fig. IV.2).

The drug discovery involves many steps to which bioinformatics can contribute. For example, possible target proteins of future drugs can be predicted from protein sequences (Han *et al.*, 2007). When ligands binding to a target protein are known, Ligand Based Virtual Screening (LBVS) methods can predict lead compounds for the target on the basis of information of chemical compounds (Zernov *et al.*, 2003; Swamidass *et al.*, 2005; Jorissen and Gilson, 2005; Hecht and Fogel, 2007; Eckert and Bajorath, 2007). With 3D structural data for the target protein available from PDB (Berman *et al.*, 2000), docking or molecular dynamics analyses can also reveal kinetic parameters (Shoichet *et al.*, 1992; Jones *et al.*, 1997; Morris *et al.*, 1998; Case *et al.*, 2005), which indicate efficacy of lead compounds. When a sufficient amount of information of kinetic parameters is available, for example, from BindingDB (Liu *et al.*, 2007), these values can be statistically predicted (Bock and Gough, 2005; Joseph *et al.*, 2008). As ligands strongly binding to the target do not always become drugs due to problems of safety, absorption, distribution, metabolism, excretion and toxicity of ligands in the body should be statistically predicted or estimated by simulation on the basis of metabolic pathways (van de Waterbeemd and Gifford, 2003; Butcher *et al.*, 2004; Ng *et al.*, 2006; Kell, 2006). In addition, as more and more information is accumulated in journals, text mining for drug discovery (Banville, 2006; Takahara *et al.*, 2008) will be more useful.

In order to realize the "bioinformatical drug discovery", first, in addition to improving our protein-chemical interaction prediction method by utilizing different datasets

including more general or more specific datasets, considering 3D structure of chemical compounds to classify chiral compounds and contriving more discriminating descriptors for both proteins and chemical compounds, we will be able to develop prediction methods dealing with other steps in drug discovery and integrate our method with these.

Acknowledgements

This thesis completes the researches that I have conducted in Sakakibara Laboratory, Department of Biosciences and Informatics, Keio University, from 2004 to 2009. Here, I express my cordial gratitudes to professors and colleagues without whose contributions to be detailed later this thesis wouldn't have been completed.

Fist of all, I thank Professor Yasubumi Sakakibara for providing me with comprehensive advice on the researches and a good environment to conduct the researches. I also thank Assistant Professor Katsuyuki Yugi, Dr. Kengo Sato, and Dr. Yasunori Osana who gave me insightful suggestions on the researches in Sakakibara Laboratory. Particularly, I would like to express my gratitude to Dr. Kengo Sato for providing me with source codes for optimizing parameters of the support vector machines utilized in Table III.6. Moreover, I express my thanks to all the colleagues in Sakakibara Laboratory, who stimulated and supported me.

In relation to the research of "Identifying cooperative transcriptional regulations using protein-protein interactions"(details are provided in Part II), I offer my thanks to Dr. Yuji Kawada for first showing me cooperativity of transcription factors.

In respect to the research of statistical prediction of interactions between a protein and a chemical compound (details are provided in Part III), I would like to express my thanks to Professor Masaya Imoto, Dr. Yusuke Minato, Mr. Hiroki Kobayashi, Mr. Takayuki Shirakawa and Mr. Kentaro Torii for experimentally verifying our computational predictions by in vitro binding assay and suggesting me the importance of utilizing additional and feedback data. In particular, I extend my thanks to Mr. Takayuki Shirakawa and Mr. Kentaro Torii for providing Fig. III.17A and Supplementary Fig. 1. My thanks also go to Dr. Takashi Komori and his team in INTEC Systems Institute, Inc. for developing the COmprehensive Predictor of Interactions between Chemical compounds And Target proteins (COPICAT) system.

Lastly, I would like to show my sincere thanks to Professor Yasubumi Sakakibara, Professor Kotaro Oka, Professor Masaya Imoto and Professor Akito Sakurai for examining and judging my doctoral dissertation.

February 2009

Nobuyoshi Nagamine

References

H.J. An, T.R. Peavy, J.L. Hedrick, and C.B. Lebrilla. Determination of N-glycosylation sites and site heterogeneity in glycoproteins. *Anal. Chem.*, 75:5628–5637, 2003.

G.D. Bader an D. Betel and C.W. Hogue. BIND: the biomolecular interaction network database. *Nucleic Acids Res.*, 31:248–250, 2003.

E. Andrianantoandro, S. Basu, D.K. Karig, and R. Weiss. Synthetic biology: new engineering rules for an emerging discipline. *Mol. Syst. Biol.*, 2:2006.0028, 2006.

R. Apweiler, A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M.J. Martin, D.A. Natale, C. O'Donovan, N. Redaschi, and L.S. Yeh. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, 32:D115–119, 2004.

N. Banerjee and M.Q. Zhang. Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res.*, 31:7024–7031, 2003.

D.L. Banville. Mining chemical structural information from the drug literature. *Drug Discov. Today*, 11:35–42, 2006.

L. Bardwell, J.G. Cook, J.X. Zhu-Shimoni, D. Voora, and J. Thorner. Differential regulation of transcription: repression by unactivated mitogen-activated protein kinase Kss1 requires the Dig1 and Dig2 proteins. *Proc. Natl. Acad. Sci. USA*, 95:15400–15405, 1998.

D. Baulcombe. DNA events. An RNA microcosm. *Science*, 297:2002–2003, 2002.

S.A. Benner and A.M. Sismour. Synthetic biology. *Nat. Rev. Genet.*, 6:533–543, 2005.

G. Berben, M. Legrain, and F. Hilger. Studies on the structure, expression and function of the yeast regulatory gene PHO2. *Gene*, 66:307–312, 1988.

H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Res.*, 28:235–242, 2000. `http://www.rcsb.org/pdb/`.

M. Bhasin and G.P. S. Raghava. GPCRpred: an SVM-based method for prediction of families and subfamilies of g-protein coupled receptors. *Nucleic Acids Res.*, 32:383–389, 2004.

P.I. Blaiseau, A.D. Isnard, Y. Surdin-Kerjan, and D. Thomas. Met31p and Met32p, two related zinc finger proteins, are involved in transcriptional regulation of yeast sulfur amino acid metabolism. *Mol. Cell. Biol.*, 17:3640–3648, 1997.

J.R. Bock and D.A. Gough. Predicting protein-protein interactinos from primary structure. *Bioinformatics*, 17:455–460, 2001.

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

N. Bouquin, A.L. Johnson, B.A. Morgan, and L.H. Johnston. Association of the cell cycle transcription factor Mbp1 with Skn7 response regulator in budding yeast. *Mol. Cell. Biol.*, 10:3389–3400, 1999.

M. Braus, O. Grundmann, S. Bruckner, and H.U. Mosch. Aminoacid starvation and Gcn4p regulate adhesive growth and FLO11 gene expression in Saccharomyces cerevisiae. *Mol. Cell. Biol.*, 14: 4272–4284, 2003.

M. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, M. Jr. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA*, 97:262–267, 2000.

C. Brun, C. Herrmann, and A. Guenoche. Clustering proteins from interaction networks for the prediction of cellular functions. *BMC Bioinformatics*, 5, 2004.

E.C. Butcher, E.L. Berg, and E.J. Kunkel. Systems biology in drug discovery. *Nat. Biotechnol.*, 22: 1253–1259, 2004.

R.A. Butcher and S.L. Schreiber. Identification of Ald6p as the target of a class of small-molecule suppressors of FK506 and their use in network dissection. *Proc. Natl. Acad. Sci. U. S. A.*, 101: 7868–7873, 2004.

P. Carninci, T. Kasukawa, S. Katayama, J. Gough, M.C. Frith, N. Maeda, R. Oyama, T. Ravasi, B. Lenhard, C. Wells, R. Kodzius, K. Shimokawa, V.B. Bajic, S.E. Brenner, S. Batalov, A.R. Forrest, M. Zavolan, M.J. Davis, L.G. Wilming, V. Aidinis, J.E. Allen, A. Ambesi-Impiombato,

R. Apweiler, R.N. Aturaliya, T.L. Bailey, M. Bansal, L. Baxter, K.W. Beisel, T. Bersano, H. Bono, A.M. Chalk, K.P. Chiu, V. Choudhary, A. Christoffels, D.R. Clutterbuck, M.L. Crowe, E. Dalla, B.P. Dalrymple, B. de Bono, G. Della Gatta, D. di Bernardo, T. Down, P. Engstrom, M. Fagiolini, G. Faulkner, C.F. Fletcher, T. Fukushima, M. Furuno, S. Futaki, M. Gariboldi, P. Georgii-Hemming, T.R. Gingeras, T. Gojobori, R.E. Green, S. Gustincich, M. Harbers, Y. Hayashi, T.K. Hensch, N. Hirokawa, D. Hill, L. Huminiecki, M. Iacono, K. Ikeo, A. Iwama, T. Ishikawa, M. Jakt, A. Kanapin, M. Katoh, Y. Kawasawa, J. Kelso, H. Kitamura, H. Kitano, G. Kollias, S.P. Krishnan, A. Kruger, S.K. Kummerfeld, I.V. Kurochkin, L.F. Lareau, D. Lazarevic, L. Lipovich, J. Liu, S. Liuni, S. McWilliam, M. Madan Babu, M. Madera, L. Marchionni, H. Matsuda, S. Matsuzawa, H. Miki, F. Mignone, S. Miyake, K. Morris, S. Mottagui-Tabar, N. Mulder, N. Nakano, H. Nakauchi, P. Ng, R. Nilsson, S. Nishiguchi, S. Nishikawa, F. Nori, O. Ohara, Y. Okazaki, V. Orlando, K.C. Pang, W.J. Pavan, G. Pavesi, G. Pesole, N. Petrovsky, S. Piazza, J. Reed, J.F. Reid, B.Z. Ring, M. Ringwald, B. Rost, Y. Ruan, S.L. Salzberg, A. Sandelin, C. Schneider, C. Schonbach, K. Sekiguchi, C.A. Semple, S. Seno, L. Sessa, Y. Sheng, Y. Shibata, H. Shimada, K. Shimada, D. Silva, B. Sinclair, S. Sperling, E. Stupka, K. Sugiura, R. Sultana, Y. Takenaka, K. Taki, K. Tammoja, S.L. Tan, S. Tang, M.S. Taylor, J. Tegner, S.A. Teichmann, H.R. Ueda, E. van Nimwegen, R. Verardo, C.L. Wei, K. Yagi, H. Yamanishi, E. Zabarovsky, S. Zhu, A. Zimmer, W. Hide, C. Bult, S.M. Grimmond, R.D. Teasdale, E.T. Liu, V. Brusic, J. Quackenbush, C. Wahlestedt, J.S. Mattick, D.A. Hume, C. Kai, D. Sasaki, Y. Tomaru, S. Fukuda, M. Kanamori-Katayama, M. Suzuki, J. Aoki, T. Arakawa, J. Iida, K. Imamura, M. Itoh, T. Kato, H. Kawaji, N. Kawagashira, T. Kawashima, M. Kojima, S. Kondo, H. Konno, K. Nakano, N. Ninomiya, T. Nishio, M. Okada, C. Plessy, K. Shibata, T. Shiraki, S. Suzuki, M. Tagami, K. Waki, A. Watahiki, Y. Okamura-Oho, H. Suzuki, J. Kawai, and Y. Hayashizaki. The transcriptional landscape of the mammalian genome. *Science*, 309:1559–1563, 2005.

D.A. Case, T.E. Cheatham, T. Darden, H. Gohlke, R. Luo, K.M. Merz, A. Onufriev, C. Simmerling, B. Wang, and R.J. Woods. The Amber biomolecular simulation programs. *J. Comput. Chem.*, 26:1668–1688, 2005.

C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

F. Chen, G.A. Rodan, and A. Schmidt. Development of selective androgen receptor modulators and their therapeutic applications. *Zhonghua Nan Ke Xue*, 8:162–168, 2002.

R.J. Cho, M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockart, and R.W. Davis. A geome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell. Biol.*, 2:65–73, 1998.

M. Clark. Generalized fragment-substructure based property prediction method. *J. Chem. Inf. Model.*, 45:30–38, 2005.

D. Compagno, C. Merle, A. Morin, C. Gilbert, J.R. Mathieu, A. Bozec, C. Mauduit, M. Benahmed, and F. Cabon. SIRNA-directed in vivo silencing of androgen receptor inhibits the growth of castration-resistant prostate carcinomas. *PLoS ONE*, 2:e1006, 2007.

C. Cortes and V. Vapnik. Support-vector network. *Machine Learning*, 20:273–297, 1995.

S. Cosconati, J.A. Hong, E. Novellino, K.S. Carroll, D.S. Goodsell, and A.J. Olson. Structure-based virtual screening and biological evaluation of Mycobacterium tuberculosis adenosine 5'-phosphosulfate reductase inhibitors. *J. Med. Chem.*, 51:6627–6630, 2008.

M.P. Cosma, T. Tanaka, and K. Nasmyth. Ordered recruitment of transcription and chromatin remodeling factors to a cell cycle- and developmentally regulated promoter. *Cell*, 97:299 – 311, 1999.

K.H. Cox, A.B. Pinchak, and T.G. Cooper. Genome-wide transcriptional analysis in S. *cerevisiae* by mini-array membrane hybridization. *Yeast*, 15:703–713, 1999.

N. Cristianini and J. Sawe-Taylor. *An Introduction to Support Vector Machines.* Cambridge University Press, Cambridge, UK, 2000.

P.S. Danielian, R. White, J.A. Less, and M.G. Parker. Identification of a conserved region required for hormone dependent transcriptional activation by steroid hormone receptors. *EMBO J.*, 11: 1025–1033, 1992.

C. Devlin, K. Tice-Baldwin, D. Shore, and K.T. Arndt. RAP1 is required for BAS1/BAS2- and GCN4-dependent transcription of the yeast HIS gene. *Mol. Cell. Biol.*, 11:3642–3651, 1991.

H. Eckert and J. Bajorath. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov. Today*, 12:225–233, 2007.

M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Bostein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.

I. Fingerman, V. Nagaraj, D. Norris, and A.K. Vershon. Sfp1 plays a key role in yeast ribosome biogenesis. *Eukaryot. Cell*, 2:1061–1068, 2003.

J. Finn, M. Stidham, M. Hilgers, and G.C. Kedar. Identification of novel inhibitors of methionyl-tRNA synthetase (MetRS) by virtual screening. *Bioorg. Med. Chem. Lett.*, 18:3932–3937, 2008.

D. Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80:27–38, 1993.

J.W. Funder and J.E. Mercer. Cimetidine, a histamine H2 receptor antagonist, occupies androgen receptors. *J. Clin. Endocrinol. Metab.*, 48:189–191, 1979.

J.M. Gancedo. Control of pseudohyphae formation in *saccharomyces cerevisiae*. *FEMS Microbiol. Rev.*, 25:107–123, 2001.

W. Gao, P.J. Reiser, C.C. Coss, M.A. Phelps, J.D. Kearbey, D.D. Miller, and J.T. Dalton. Selective androgen receptor modulator treatment improves muscle strength and body composition and prevents bone loss in orchidectomized rats. *Endocrinology*, 146:4887–4897, 2005.

J. Gasteiger and T. Engel. *Chemoinformatics: A Textbook*. Wiley-VCH, Weinheim, 2003.

S.M. Gomez, W.S. Noble, and A. Rzhetsky. Learning to predict protein-protein interactions. *Bioinformatics*, 19:1875–1881, 2003.

R. Guha, M.T. Howard, G.R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. Wegner, and E.L. Willighagen. The Blue Obelisk-interoperability in chemical informatics. *J. Chem. Inf. Model.*, 46:991–998, 2006.

S. Gunther, M. Kuhn, M. Dunkel, M. Campillos, C. Senger, E. Petsalaki, J. Ahmed, E.G. Urdiales, A. Gewiess, L.J. Jensen, R. Schneider, R. Skoblo, R.B. Russell, P.E. Bourne, P. Bork, and R. Preissner. SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res.*, 36:D919–922, 2008.

L.Y. Han, C.J. Zheng, B. Xie, J. Jia, X.H. Ma, F. Zhu, H.H. Lin, X. Chen, and Y.Z. Chen. Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness. *Drug Discov. Today*, 12:304–313, 2007.

R. Hasan, C. Leroy, A.D. Isnard, J. Labarre, E. Boy-Marcotte, and M.B. Toledano. The control of the yeast h_2o_2 response by the Msn2/4 transcription factors. *Mol. Microbiol.*, 45:233–241, 2002.

D. Hecht and G.B. Fogel. High-throughput ligand screening via preclustering and evolved neural networks. *IEEE/ACM Trans Comput Biol Bioinform*, 4:476–484, 2007.

S. Hermann-LeDenmat, M. Werner, A. Sentenac, and P. Thuriaw. Suppression of yeast RNA polymerase 3 mutations by FHL1, a gene coding for a fork head protein involved in rRNA processing. *Mol. Cell. Biol.*, 14:2905–2913, 1994.

Y. Ho, M. Costanzo, L. Moore, R. Kobayashi, and B.J. Andrews. Regulation of transcription at the Saccharomyces cerevisiae start transition by Stb1 a swi6-binding protein. *Mol. Cell. Biol.*, 19: 5267–5278, 1999.

M. Huynen, B. Snel, W. Lathe, and P. Bork. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.*, 10:1204–1210, 2000.

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

A. Jamai, E. Dubois, A.K. Vershon, and F. Messenguy. Swapping functional specificity of a MADS Box protein: resudues required for Arg80 regulation of arginine metabolism. *Mol. Cell. Biol.*, 22: 5741–5752, 2002.

R. Jansen, D. Greenbaum, and M. Gerstein. Relating whole-genome expression data with protein-protein interactions. *Genome Res.*, 12:37–46, 2002.

R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N.J. Krogan, S. Chung, A. Emili, M. Snyder, J.F. Greenblatt, and M. Gerstein. A Baysian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302:449–453, 2003.

E. Jenwitheesuk and R. Samudrala. Prediction of HIV-1 protease inhibitor resistance using a protein-inhibitor flexible docking approach. *Antivir. Ther. (Lond.)*, 10:157–166, 2005.

G. Jones, P. WIllett, R.C. Glen, A.R. Leach, and R. Taylor. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, 267:727–748, 1997.

R.N. Jorissen and M.K. Gilson. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.*, 45:549–561, 2005.

T.B. Joseph, B.V. Suneel Kumar, B. Santhosh, S. Kriti, A.B. Pramod, M. Ravikumar, and M. Kishore. Quantitative structure activity relationship and pharmacophore studies of adenosine receptor a inhibitors. *Chem Biol Drug Des*, 72:395–408, 2008.

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

R. Karchin, K. Karplus, and D. Haussler. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics*, 18:147–159, 2002.

D.B. Kell. Systems biology, metabolic modelling and metabolomics in drug discovery and development. *Drug Discov. Today*, 11:1085–1092, 2006.

I. Kinoyama, N. Taniguchi, A. Toyoshima, E. Nozawa, T. Kamikubo, M. Imamura, M. Matsuhisa, K. Samizu, E. Kawanimani, T. Niimi, N. Hamada, H. Koutok, T. Furutani, M. Kudoh, M. Okada, M. Ohta, and S. Tsukamoto. (+)-(2R,5S)-4-[4-cyano-3-(trifluoromethyl)phenyl]-2,5-dimethyl-N-[6-(trifluoromethyl)pyridin-3-yl]piperazine-1-carboxamide (YM580) as an orally potent peripherally selective nonsteroidal androgen receptor antagonist. *J. Med. Chem.*, 49:716–726, 2006.

H. Kitano. Computational systems biology. *Nature*, 420:206–210, 2002.

T. Klabunde and G. Hessler. Drug design strategies for targeting G-protein-coupled receptors. *Chembiochem*, 3:928–944, 2002.

C. Koch, T. Moll, M. Neuberg, H. Ahorn, and K. Nasmyth. A role for the transcription factors Mbp1 and Swi4 in progression from G_1 to S phase. *Science*, 261:1551–1557, 1993.

M. Koranda, A. Shleiffer, L. Endler, and G. Ammerer. Forkhead-like transcription factors result Ndd1 to the chromatin of G_2/M-specific promoters. *Nature*, 406:94–98, 2000.

K. Kristiansen. Molecular mechanisms of ligand binding, signaling and regulation within G-protein-coupled receptors:molecular modeling and mutagenesis approaches to receptor structures and function. *Pharmacol. Ther.*, 103:21–80, 2004.

R. Kumar, D.M. Reynolds, A. Shevechenko, A. Schevchenko, S.D. Goldstone, and S. Dalton. Forkhead transcription factors, Fkh1p and Fkh2p, collaborate with Mcm1p to control transcription required for M-phase. *Curr. Biol.*, 10:898–896, 2000.

J. Lee, C. Godon, G. Lagniel, D. Spector, J. Garin, J. Labarre, and M.B. Toledano. Yap1 and Skn7 control two specializad oxidative stress response regulons in yeast. *J. Biol. Chem.*, 274: 16040–16046, 1999.

R.C. Lee, R.L. Feinbaum, and V. Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75:843–854, 1993.

T.I. Lee, N.J. Rinaldi, F. Robert, D.T. Odom, Z. Bar-Joseph, G.K. Gerber, N.M. Hannett, C.T. Harbison, C.M. Thompson, I. Simon, J. Zeitlinger, E.G. Jennings, H.L. Murray, D.B. Gordon, B. Ren, J.J. Wyrick, J.B. Tagne, T.L. Volkert, E. Fraenkel, D.K. Gifford, and R.A. Young. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science*, 298:799–804, 2002.

C.S. Leslie, E. Eskin, A. Cohen, J. Weston, and W.S. Noble. Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20:467–476, 2004.

L. Liao and S. Noble. Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *J. Comput. Biol.*, 10:857–868, 2003.

J.T. Link, B. Sorensen, J. Patel, M. Grynfarb, A. Goos-Nilsson, J. Wang, S. Fung, D. Wilcox, B. Zinker, P. Nguyen, B. Hickman, J.M. Schmidt, S. Swanson, Z. Tian, T.J. Reisch, G. Rotert, J. Du, B. Lane, T.W. von Geldern, and P.B. Jacobson. Antidiabetic activity of passive nonsteroidal glucocorticoid receptor modulators. *J. Med. Chem.*, 48:5295–5304, 2005.

K. Liolios, K. Mavromatis, N. Tavernarakis, and N.C. Kyrpides. The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.*, 36:D475–479, 2008.

T. Liu, Y. Lin, X. Wen, R.N. Jorissen, and M.K. Gilson. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, 35:198–201, 2007.

C.J. Loy, D. Ldall, and U. Surana. Ndd1, a high-dosage suppressor of cdc28-1N, is essential for

expression of a subset of late-S-phase-specific gene transcription in S. *cerevisiae*. *Mol. Cell. Biol.*, 19:3312–3327, 1999.

Y.M. Mamnun, R. Pandjaitan, Y. Mahc, A. Delahodde, and K. Kuchler. The yeast zinc finger regulators Pdr1pand Pdr3p control pleiotropic drug resistance (PDR) as homo- and heterodimers *in vivo*. *Mol. Microbiol.*, 46:1429–1440, 2002.

L.L. Marchand, J.H. Hankin, L.R. Wilkens, L.M. Pierre, A. Franke, L.N. Kolonel, A. Seifried, L.J. Custer, W. Chang, A. Lum−Jones, and T. Donlon. Combined effects of welldone red meat and smoking and rapid N−acetyltransferase 2 and CYPIA2 phenotypes in increasing colorectal cancer risk. *Cancer Epideminology, Biomarkers & Prevention*, 10:1259–1266, 2001.

S. Martin, D. Roe, and J.-L. Faulon. Predicting protein-protein interactions using signature products. *Bioinformatics*, 21:218–226, 2005.

D.S. McNabb, Y. Xing, and L. Guarente. Cloning of yeast HAP5 : a novel subunit of a heterotrimeric complex required for CCAAT binding. *Genes Dev.*, 9:47–58, 1995.

C. Merlot, D. Domine, C. Cleva, and D.J. Church. Chemical substructures in drug discovery. *Drug Discov. Today*, 8:594–602, 2003.

H.W. Mewes, U. Guldener, K. Mayer, M. Mokrejs, B. Morgenstern, M. Munsterkotter, S. Rudd, and B. Weil. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, 30:31–34, 2002.

K. Miyoshi, T. Miyakawa, and K. Mizuta. Repression of rRNA synthesis due to a secretory defect requires the C-terminal silencing domain of Rap1p in Saccharomyces cerevisiae. *Nucleic Acids*, 29:3297–3303, 2001.

M. Mollapour, D. Fong, K. Balakrishnan, N. Harris, S. Thompson, C. Schuller, K. Kuchler, and P.W. Piper. Screeningn the yeast deletant mutatnt collection for hypersensitivity and hyper-resistance to sorbate, a weak organic acid food preservative. *Yeast*, 21:927–946, 2004.

B.A. Morgan, G.R. Banks, W.M. Toone, D. Raitt, S. Kuge, and L.H. Johnston. The Skn7 response regulator controls gene expression in the oxidative stress response of the budding yeast *saccharomyces cerevisiae*. *EMBO J.*, 16:1035–1044, 1997.

G.M. Morris, D.S. Goodsell, R.S. Halliday, R. Huey, W.E. Hart, R.K. Belew, and A.J. Olson. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.*, 19:1639–1662, 1998.

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

N. Nagamine, Y. Kawada, and Y. Sakakibara. Identifying cooperative transcriptional regulations using protein-protein interactions. *Nucleic Acids Res.*, 33:4828–4837, 2005.

A. Ng, B. Bursteinas, Q. Gao, E. Mollison, and M. Zvelebil. Resources for integrative systems biology: from data through databases to networks and dynamic system models. *Brief. Bioinformatics*, 7: 318–330, 2006.

Y. Okuno, A. Tamon, H. Yabuuchi, S. Niijima, K. Tomoura, and C. Feng. GLIDA: GPCR–ligand database for chemical genomics drug discovery and tools update. *Nucleic Acids Res.*, 36:D907–912, 2008.

K. Palczewski, T. Kumasaka, T. Hori, C.A. Behnke, H. Motoshima, B.A. Fox, I. Le Trong, D.C. Teller, T. Okada, R.E. Stenkamp, M. Yamamoto, and M. Miyano. Crystal structure of rhodopsin: A G protein-coupled receptor. *Science*, 289:739–745, 2000.

K. Pereira de Jesus-Tran, P.L. Cote, L. Cantin, J. Blanchet, F. Labrie, and R. Breton. Comparison of crystal structures of human androgen receptor ligand-binding domain complexed with various agonists reveals molecular determinants responsible for binding affinity. *Protein Sci.*, 15:987–999, 2006.

A. Pic, F.L. Lim, S.J. Ross, E.A. Vcal, A.L. Johnson, M.R. Sultan, A.G. West, L.H. Johnston, A.D. Sharrocks, and B.A. Morgan. The forkhead protein Fkh2 is a component of the yeast cell cycle transcription factor SFF. *EMBO J.*, 19:3750–3761, 2000.

J. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In A. Smola, P. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, Cambridge, MA, USA, 2000.

S.G. Rasmussen, H.J. Choi, D.M. Rosenbaum, T.S. Kobilka, F.S. Thian, P.C. Edwards, M. Burghammer, V.R. Ratnala, R. Sanishvili, R.F. Fischetti, G.F. Schertler, W.I. Weis, and B.K. Kobilka.

Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. *Nature*, 450:383–387, 2007.

L. Riego, A. Avendano, A. DeLuna, E. Rodriguez, and A. Gonzalerz. GDH1 expression is regulated by GLN3, GCN4, and HAP4 under respiratory growth. *Biochem. Biophy. Res. Commun.*, 293: 79–85, 2002.

B. D Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, UK, 1996.

C.E. Roselli. The effect of anabolic-androgenic steroids on aromatase activity and androgen receptor binding in the rat preoptic area. *Brain Res.*, 792:271–276, 1998.

B.L. Roth, E. Lopez, S. Patel, and W.K. Kroeze. The Multiplicity of Serotonin Receptors: Uselessly Diverse Molecules or an Embarrassment of Riches? *The Neuroscientist*, 6:252–262, 2000.

G. Rudnick and S.C. Wall. The molecular mechanism of {ecstasy} [3,4-methylenedioxymethamphetamine, mdma]: serotonin transporters are targets for mdma induced serotonin release. *Proc. Natl. Acad. Sci. USA*, 89:1817–1821, 1992.

A.P. Russ and S. Lampel. The druggable genome: an update. *Drug Discov. Today*, 10:1607–1610, 2005.

J. Sadowski and J. Gasteiger. From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. *Chemical Reviews*, 93:2567–2581, 1993.

B. Scholkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, and R.C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 2001.

I. Schomburg, A. Chang, C. Ebeling, M. Gremse, C. Heldt, G. Huhn, and D. Schomburg. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.*, 32:D431–433, 2004.

H.J. Schuller, A. Schutz, S. Knab, B. Hoffmann, and E. Schweizer. Importance of general regulatory factos Rap1p, Abf1p and Reb1p for the activation of yeast fatty acid synthease genes FAS1 and FAS2. *Eur. J. Biochem.*, 225:213–222, 1994.

B. Schwikowski, P. Uetz, and S. Fields. A network of protein-protein interactions in yeast. *Nat. Biotechnol.*, 18:1257–1261, 2000.

B.K. Shoichet, D.L. Bodian, and I.D. Kuntz. Molecular docking using shape descriptors. *J. Comput. Chem.*, 13:380–397, 1992.

I. Simon, J. Barnett, N. Hanett, C.T. Harbison, N.J. Rinaldi, T.L. Volkert, J.J. Wyrick, J. Zeitlinger, D.K. Gifford, T.S. Jaakkola, and R.A. Young. Serial regulation of transcriptional regulators in the yeast cell cycle. *Science*, 106:697–708, 2001.

T. Soga, Y. Ueno, H. Naraoka, Y. Ohashi, M. Tomita, and T Nishioka. Simultaneous determination of anionic intermediates for *bacillus subtilis* metabolic pathways by capillary electrophoresis electrospray ionization mass spectrometry. *Anal. Chem.*, 74:2233–2239, 2002.

M.S. Spector, A. Raff, Heshani. DeSilva, K. Lee, and M.A. Osley. Hir1p and Hir2p function as transcriptional corepressors to regulate histone gene transcription in the S. *cerevisiae* cell cycle. *Mol. cell. Biol.*, 17:545–552, 1997.

J.E. Sprague, M.L. Banks, V.J. Cook, and E.M. Mills. Hypothalamic-pituitary-thyroid axis and sympathetic nervous system involvement in hyperthermia induced by 3,4-methylenedioxymethamphetamine (Ecstasy). *J. Pharmacol. Exp. Ther.*, 305:159–166, 2003.

E. Sprinzak and H. Margalit. Correlated sequence-signatures as markers of protein-protein interaction. *J. Mol. Biol.*, 311:681–692, 2001.

G. Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16:16–23, 2000.

S.J. Swamidass, J. Chen, J. Bruand, P. Phung, L. Ralaivola, and P. Baldi. Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinformatics*, 21:359–368, 2005.

R.J. Taft, M. Pheasant, and J.S. Mattick. The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays*, 29:288–299, 2007.

Y. Takahara, T. Kobayashi, K. Takemoto, T. Adachi, K. Osaki, K. Kawahara, and G. Tsujimoto. Pharmacogenomics of cardiovascular pharmacology: development of an informatics system for analysis of DNA microarray data with a focus on lipid metabolism. *J. Pharmacol. Sci.*, 107:1–7, 2008.

A.E. Teschendorff, Y. Wang, N.L. Barbosa-Morais, J.D. Brenton, and C. Caldas. A variational

Bayesian mixture modelling framework for cluster analysis of gene-expression data. *Bioinformatics*, 21:3025–3033, 2005.

I.V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl, V.A. Palyulin, E.V. Radchenko, N.S. Zefirov, A.S. Makarenko, V.Y. Tanchuk, and V.V. Prokopenko. Virtual computational chemistry laboratory–design and description. *J. Comput. Aided Mol. Des.*, 19:453–463, 2005.

S. Turkel and L.F. Bisson. Transcription of the HXT4 gene is regulated by Gcr1p and Gcr2p in the yeast S. *cerevisiae*. *Yeast*, 15:1045–1097, 1999.

M.S. Tuttle, D. Radisky, L. Li, and J. Kaplan. A dominant allele of PDR1 alters transition metal resistance in yeast. *J. Biol. Chem.*, 278:1273–1280, 2003.

H. Uemura, M. Koshio, Y. Inoue, M.C. Lopez, and H.V. Baker. The role of Gcr1p in the transcriptional activation of glycolytic genes in yeast Saccharomyces cerevisiae. *Genetics*, 147:521–532, 1997.

H. van de Waterbeemd and E. Gifford. ADMET in silico modelling: towards prediction paradise? *Nat Rev. Drug Discov.*, 2:192–204, 2003.

V. Vapnik. *Statistical Learning Theory.* Wiley, New York, USA, 1998.

M.S. Venkatarajan and W. Braun. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties. *Journal of Molecular Modeling*, 7:445–453, 2001.

V.K. Vyas, S. Kuchin, C.D. Berkey, and M. Carlson. Snf1 kinases with different beta-subunit isoforms play distinct roles in regulation haploid invasive growth. *Mol. Cell. Biol.*, 23:1341–1348, 2003.

C. Wang, C. Ding, R.F. Meraz, and S.R. Holbrook. PSoL: a positive sample only learning algorithm for finding non-coding RNA genes. *Bioinformatics*, 22:2590–2596, 2006.

G.L. Wheeler, E.W. Trotter, I.W. Dawes, and C.M. Grant. Coupling of the transcriptional regulation of glutathione biosynthesis to the availability of glutathione and methidonine via the met4 and yap1 transcription factors. *J. Biol. Chem.*, 278:49920–49928, 2003.

G. Wiederrecht, D. Seto, and C.S. Parker. Isolation of the gene encoding the S. *cerevisiae* heat shock transcription factor. *Cell*, 54:841–853, 1988.

D.S. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, pages D901–D906, 2008.

I. Xenarios, L. Salwinski, X.J. Duan, P. Higney, S.M. Kim, and D. Eisenberg. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acis Res.*, 30:303–305, 2002.

X. Xiao, S.H. Shao, Z.D. Huang, and K.C. Chou. Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. *J. Comput. Chem.*, 27: 478–482, 2006.

Y. Xue, Z.R. Li, C.W. Yap, L.Z. Sun, X. Chen, and Y.Z. Chen. Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *J. Chem. Inf. Comput. Sci.*, 44:1630–1638, 2004.

Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, pages i232–240, 2008.

C. Yanover and T. Hertz. Predicting protein-peptide binding affinity by learning peptide-peptide distance functions. In *RECOMB 2005*, pages 456–471, 2005.

M.A. Yildirim, K.I. Goh, M.E. Cusick, A.L. Barabasi, and M. Vidal. Drug-target network. *Nat. Biotechnol.*, 25:1119–1126, 2007.

C.S. Yu, Y.C. Chen, C.H. Lu, and J.K. Hwang. Prediction of protein subcellular localization. *Proteins*, 64:643–651, 2006.

H. Yu, X. Zhu, D. Greenbaum, J. Karro, and M. Gerstein. TopNet: a tool for comparing biological subnetworks, correlating protein properties with topological statistics. *Nucleic Acids Res.*, 32, 2004.

J.H. Zar. *Biostatistical analysis (4th edition).* Prentice Hall International, Upper Saddle River, NJ, 1998.

V.V. Zernov, K.V. Balakin, A.A. Ivaschenko, N.P. Savchuk, and I.V. Pletnev. Drug discovery using

support vector machines. the case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J. Chem. Inf. Comput. Sci.*, 43:2048–2056, 2003.

S. Zhong, X. Chen, X. Zhu, B. Dziegielewska, K.E. Bachman, T. Ellenberger, J.D. Ballin, G.M. Wilson, A.E. Tomkinson, and A.D. MacKerell. Identification and validation of human DNA ligase inhibitors using computer-aided drug design. *J. Med. Chem.*, 51:4553–4562, 2008.

G. Zhu, P.T. Spellman, T. Volpe, P.O. Brown, D. Botstein, T.N. Davis, and B. Futcher. Two yeast forkhead genes regulate the cell cycle and pseudohyphal growth. *Nature*, 406:90–94, 2000.

Z. Zhu, R.R. Becklin, D.M. Desideio, and J.T. Dalton. Mass spectrometric characterization of the human androgen receptor ligand-binding domain expressed in Escherichia coli. *Biochemistry*, 40: 10756–10763, 2001.

Index

Appendix A - Supplementary Tables and Figures

Supplementary Table A1 $P_B < 0.001$, $O_{\min} = 2$, $I_{\min} = 3$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	HIR1	HIR2	5.41E-07	4.14E-07	Loy *et al.*, 1999
2	FHL1	RAP1	3.48E-04	8.10E-34	NA
3	MCM1	SWI4	6.53E-04	2.57E-04	NA
4	RAP1	YAP5	1.41E-03	1.43E-10	NA
5	FKH1	NDD1	2.09E-03	9.63E-04	Kumar *et al.*, 2000
6	FHL1	PDR1	2.29E-03	2.30E-06	NA
7	FKH2	MCM1	3.99E-03	3.02E-04	Spector *et al.*, 1997
8	MBP1	MSN4	4.23E-03	2.88E-03	NA
9	ARG80	ARG81	5.61E-03	2.92E-04	Mamnun *et al.*, 2002
10	NRG1	PHD1	6.35E-03	8.02E-05	NA
11	RAP1	SFP1	6.42E-03	1.15E-03	NA
12	FHL1	GAT3	6.66E-03	3.51E-08	NA
13	FHL1	YAP5	7.52E-03	6.77E-18	NA
14	HAP2	HAP5	2.51E-04	1.01E-02	McNabb *et al.*, 1995
15	HAP3	HAP5	7.96E-04	1.01E-02	McNabb *et al.*, 1995
16	NRG1	YAP6	4.58E-07	1.17E-02	NA
17	MBP1	MCM1	1.17E-02	2.31E-04	NA
18	FKH1	FKH2	3.88E-03	1.43E-02	Koranda *et al.*, 2000; Spector *et al.*, 1997
19	HSF1	RAP1	1.44E-02	1.97E-02	NA
20	SWI5	YAP5	2.25E-02	1.17E-02	NA
21	NDD1	SKN7	3.09E-02	1.42E-02	NA
22	MBP1	SKN7	3.59E-02	1.80E-05	Bouquin *et al.*, 1999
23	GCR1	GCR2	4.21E-02	2.41E-02	Turkel and Bisson, 1999; Uemura *et al.*, 1997
24	SKN7	YAP1	8.15E-03	4.34E-02	Lee *et al.*, 1999

Supplementary Table A2 $P_B < 0.001$, $O_{\min} = 5$, $I_{\min} = 3$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	HIR1	HIR2	5.41E-07	4.14E-07	Loy *et al.*, 1999
2	FHL1	RAP1	3.48E-04	8.10E-34	NA
3	MCM1	SWI4	6.53E-04	2.57E-04	NA
4	RAP1	YAP5	1.41E-03	1.43E-10	NA
5	FKH1	NDD1	2.09E-03	9.63E-04	Kumar *et al.*, 2000
6	FHL1	PDR1	2.29E-03	2.30E-06	NA
7	FKH2	MCM1	3.99E-03	3.02E-04	Spector *et al.*, 1997
8	ARG80	ARG81	5.61E-03	2.92E-04	Mamnun *et al.*, 2002
9	NRG1	PHD1	6.35E-03	8.02E-05	NA
10	RAP1	SFP1	6.42E-03	1.15E-03	NA
11	FHL1	GAT3	6.66E-03	3.51E-08	NA
12	FHL1	YAP5	7.52E-03	6.77E-18	NA
13	HAP2	HAP5	2.51E-04	1.01E-02	McNabb *et al.*, 1995
14	HAP3	HAP5	7.96E-04	1.01E-02	McNabb *et al.*, 1995
15	NRG1	YAP6	4.58E-07	1.17E-02	NA
16	MBP1	MCM1	1.17E-02	2.31E-04	NA
17	FKH1	FKH2	3.88E-03	1.43E-02	Koranda *et al.*, 2000; Spector *et al.*, 1997
18	SWI5	YAP5	2.25E-02	1.17E-02	NA
19	NDD1	SKN7	3.09E-02	1.42E-02	NA
20	MBP1	SKN7	3.59E-02	1.80E-05	Bouquin *et al.*, 1999
21	GCR1	GCR2	4.21E-02	2.41E-02	Turkel and Bisson, 1999; Uemura *et al.*, 1997
22	SKN7	YAP1	8.15E-03	4.34E-02	Lee *et al.*, 1999

Supplementary Table A3 $P_B < 0.001$, $O_{\min} = 10$, $I_{\min} = 3$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	FHL1	RAP1	3.48E-04	8.10E-34	NA
2	MCM1	SWI4	6.53E-04	2.57E-04	NA
3	RAP1	YAP5	1.41E-03	1.43E-10	NA
4	FKH1	NDD1	2.09E-03	9.63E-04	Kumar *et al.*, 2000
5	FHL1	PDR1	2.29E-03	2.30E-06	NA
6	FKH2	MCM1	3.99E-03	3.02E-04	Spector *et al.*, 1997
7	NRG1	PHD1	6.35E-03	8.02E-05	NA
8	FHL1	YAP5	7.52E-03	6.77E-18	NA
9	NRG1	YAP6	4.58E-07	1.17E-02	NA
10	MBP1	MCM1	1.17E-02	2.31E-04	NA
11	FKH1	FKH2	3.88E-03	1.43E-02	Koranda *et al.*, 2000; Spector *et al.*, 1997
12	SWI5	YAP5	2.25E-02	1.17E-02	NA
13	NDD1	SKN7	3.09E-02	1.42E-02	NA
14	MBP1	SKN7	3.59E-02	1.80E-05	Bouquin *et al.*, 1999

Supplementary Table A4 $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 1$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	HIR1	HIR2	3.85E-09	9.37E-10	Loy *et al.*, 1999
2	ACE2	FKH2	2.94E-10	1.54E-08	Pic *et al.*, 2000
3	MBP1	SKN7	3.51E-08	9.85E-12	Bouquin *et al.*, 1999
4	NRG1	PHD1	1.19E-12	2.72E-07	NA
5	NDD1	SWI4	2.23E-12	2.85E-07	NA
6	SWI4	SWI5	6.99E-07	1.14E-09	Cosma *et al.*, 1999
7	FHL1	PDR1	9.89E-06	2.90E-22	NA
8	NDD1	RME1	2.42E-06	1.00E-05	NA
9	HAP3	HAP5	5.34E-05	4.10E-05	McNabb *et al.*, 1995
10	HAP2	HAP5	1.50E-04	4.10E-05	McNabb *et al.*, 1995
11	PHD1	SKN7	3.49E-09	1.78E-04	NA
12	FKH2	SWI5	2.08E-04	1.03E-06	Pic *et al.*, 2000
13	MBP1	MSN4	2.24E-04	5.56E-05	NA
14	FKH2	SWI4	2.35E-04	1.13E-14	NA
15	MBP1	RME1	3.36E-04	1.00E-05	NA
16	PDR1	RAP1	6.40E-13	5.43E-04	NA
17	ACE2	PHO4	3.43E-04	5.72E-04	NA
18	MBP1	PHO4	3.35E-04	5.72E-04	NA
19	PHO4	SKN7	5.72E-04	7.00E-05	NA
20	FHL1	GAT3	6.57E-04	4.93E-11	NA
21	HAP2	HAP3	9.08E-04	4.06E-04	McNabb *et al.*, 1995
22	NDD1	YAP1	3.38E-08	1.91E-03	NA
23	MBP1	YAP1	2.04E-04	1.91E-03	NA
24	ACE2	YAP1	2.18E-04	2.17E-03	NA
25	SWI5	YAP1	1.53E-05	2.17E-03	NA
26	RME1	SWI4	5.83E-03	3.15E-03	NA
27	RME1	SWI5	5.83E-03	1.68E-03	NA
28	SKN7	SWI5	6.15E-03	6.74E-05	NA
29	MAC1	ROX1	6.32E-03	4.93E-04	NA
30	ABF1	REB1	6.33E-03	4.41E-03	Schuller *et al.*, 1994
31	ACE2	SWI5	6.45E-03	2.23E-27	Zhu *et al.*, 2000
32	RME1	SKN7	8.18E-03	2.61E-03	NA
33	ACE2	RME1	9.93E-03	5.83E-03	NA
34	FHL1	RAP1	1.03E-02	7.25E-165	NA
35	FKH2	YAP1	1.17E-03	1.07E-02	NA
36	FKH2	RME1	1.31E-02	5.83E-03	NA
37	CIN5	GCN4	1.67E-04	1.32E-02	NA
38	GCN4	RAP1	1.20E-02	1.40E-02	Devlin *et al.*, 1991
39	FKH2	PHO4	1.31E-02	1.49E-02	NA
40	NDD1	PHO4	1.44E-03	1.49E-02	NA
41	PHO4	RME1	1.49E-02	5.83E-03	NA
42	PHO4	SWI4	1.49E-02	3.15E-03	NA
43	PHO4	SWI5	1.49E-02	1.68E-03	NA
44	SKN7	YAP6	1.55E-02	1.54E-03	NA
45	GCR1	GCR2	1.94E-02	4.12E-03	Turkel and Bisson, 1999; Uemura *et al.*, 1997
46	HAP2	HAP4	2.75E-02	1.28E-02	Riego *et al.*, 2002
47	CBF1	MET4	7.49E-03	2.96E-02	Wheeler *et al.*, 2003
48	PHO4	YAP1	1.49E-02	4.23E-02	NA
49	RME1	YAP1	5.83E-03	4.23E-02	NA
50	SWI4	YAP1	3.15E-03	4.23E-02	NA
51	FHL1	YAP5	4.46E-02	5.72E-72	NA

Supplementary Table A5 $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 2$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	FHL1	RAP1	3.42E-10	1.01E-60	NA
2	HIR1	HIR2	6.85E-08	5.99E-09	Loy *et al.*, 1999
3	MCM1	SWI4	2.11E-06	1.78E-05	NA
4	ACE2	FKH2	6.91E-07	3.25E-05	Pic *et al.*, 2000
5	MBP1	SKN7	2.10E-04	6.10E-10	Bouquin *et al.*, 1999
6	FKH2	MCM1	3.08E-04	4.19E-12	Spector *et al.*, 1997
7	FHL1	PDR1	4.48E-04	1.76E-12	NA
8	HAP2	HAP5	1.54E-04	5.69E-04	McNabb *et al.*, 1995
9	HAP3	HAP5	2.45E-04	5.69E-04	McNabb *et al.*, 1995
10	MBP1	MSN4	1.12E-03	1.37E-03	NA
11	NRG1	PHD1	1.90E-03	2.62E-04	NA
12	FKH2	MBP1	3.21E-04	2.05E-03	NA
13	NDD1	RME1	3.05E-03	2.34E-03	NA
14	FHL1	GAT3	3.15E-03	9.40E-09	NA
15	SWI5	YAP1	3.77E-03	3.67E-03	NA
16	NDD1	YAP1	7.70E-04	4.28E-03	NA
17	ACE2	YAP1	4.41E-03	3.67E-03	NA
18	MCM1	SWI6	1.37E-04	4.53E-03	NA
19	MBP1	MCM1	4.58E-03	6.43E-06	NA
20	NRG1	YAP6	6.49E-07	5.40E-03	NA
21	MAC1	ROX1	7.22E-03	1.12E-06	NA
22	MBP1	YAP1	7.94E-03	3.23E-03	NA
23	NDD1	SKN7	8.78E-03	4.45E-03	NA
24	SWI5	YAP5	1.08E-02	7.51E-03	NA
25	HAP2	HAP3	1.58E-02	1.35E-02	McNabb *et al.*, 1995
26	ACE2	SKN7	1.64E-02	1.96E-05	NA
27	RAP1	YAP5	1.68E-02	1.64E-13	NA
28	FKH2	SWI5	2.00E-02	6.92E-03	Pic *et al.*, 2000
29	NRG1	RLM1	2.64E-02	2.00E-02	NA
30	ABF1	REB1	2.84E-02	2.17E-02	Schuller *et al.*, 1994
31	ACE2	NDD1	3.51E-02	5.48E-03	NA
32	GCR1	GCR2	4.21E-02	8.72E-03	Turkel and Bisson, 1999; Uemura *et al.*, 1997
33	CIN5	RAP1	3.06E-03	4.65E-02	NA

Supplementary Table A6 $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 4$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	PDR1	RAP1	6.43E-11	1.10E-09	NA
2	FHL1	PDR1	1.42E-07	5.53E-10	NA
3	HIR1	HIR2	1.94E-06	3.36E-06	Loy *et al.*, 1999
4	RAP1	YAP5	5.99E-09	3.53E-05	NA
5	ABF1	REB1	1.02E-04	4.91E-05	Schuller *et al.*, 1994
6	ARG80	ARG81	5.61E-03	2.92E-04	Mamnun *et al.*, 2002
7	HAP2	HAP5	2.51E-04	1.01E-02	McNabb *et al.*, 1995
8	HAP3	HAP5	7.96E-04	1.01E-02	McNabb *et al.*, 1995
9	MBP1	MSN4	1.02E-02	7.03E-03	NA
10	RAP1	SFP1	1.45E-02	1.98E-03	NA
11	MET31	MET4	5.96E-03	2.10E-02	Blaiseau *et al.*, 1997
12	HSF1	RAP1	1.90E-02	2.90E-02	NA
13	FKH1	NDD1	2.96E-02	5.26E-03	Kumar *et al.*, 2000
14	PDR1	YAP5	3.15E-02	1.70E-03	NA
15	MCM1	SWI4	3.25E-02	4.32E-03	NA

Supplementary Table A7 $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 5$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	PDR1	RAP1	3.41E-09	2.30E-09	NA
2	FHL1	PDR1	2.27E-07	1.00E-06	NA
3	HIR1	HIR2	8.40E-06	4.10E-05	Loy *et al.*, 1999
4	RAP1	YAP5	1.01E-07	7.04E-05	NA
5	ABF1	REB1	1.25E-04	5.09E-05	Schuller *et al.*, 1994
6	FKH2	NDD1	1.25E-04	2.47E-06	Kumar *et al.*, 2000
7	ACE2	NDD1	8.62E-04	1.16E-04	NA
8	MBP1	PHO4	1.37E-03	1.44E-03	NA
9	PHO4	SKN7	1.44E-03	1.46E-04	NA
10	ACE2	PHO4	1.64E-03	1.44E-03	NA
11	NDD1	SWI4	2.49E-04	2.27E-03	NA
12	ACE2	YAP1	4.48E-04	4.62E-03	NA
13	MBP1	YAP1	4.14E-04	4.62E-03	NA
14	SKN7	YAP1	5.38E-05	4.62E-03	Lee *et al.*, 1999
15	SWI5	YAP1	1.22E-04	4.62E-03	NA
16	RME1	SKN7	5.01E-03	3.88E-03	NA
17	FKH2	RME1	5.09E-03	5.01E-03	NA
18	ARG80	ARG81	5.61E-03	2.92E-04	Mamnun *et al.*, 2002
19	FKH2	YAP1	1.49E-04	5.69E-03	NA
20	RME1	SWI5	5.01E-03	5.98E-03	NA
21	RME1	SWI4	5.01E-03	6.01E-03	NA
22	NDD1	SWI5	8.27E-03	7.06E-03	NA
23	HAP2	HAP5	2.51E-04	1.01E-02	McNabb *et al.*, 1995
24	HAP3	HAP5	1.80E-03	1.01E-02	McNabb *et al.*, 1995
25	FKH2	PHO4	5.09E-03	1.22E-02	NA
26	NDD1	PHO4	3.35E-03	1.22E-02	NA
27	PHO4	RME1	1.22E-02	5.01E-03	NA
28	PHO4	SWI4	1.22E-02	6.01E-03	NA
29	PHO4	SWI5	1.22E-02	5.98E-03	NA
30	ACE2	RME1	1.29E-02	5.01E-03	NA
31	RAP1	SFP1	1.76E-02	1.98E-03	NA
32	ACE2	MBP1	3.74E-03	1.84E-02	NA
33	MBP1	MSN4	1.92E-02	2.02E-02	NA
34	MBP1	SKN7	2.23E-02	6.89E-05	Bouquin *et al.*, 1999
35	ACE2	FKH2	1.65E-02	2.25E-02	Pic *et al.*, 2000
36	FKH2	SWI5	3.51E-02	1.99E-02	Pic *et al.*, 2000
37	PHO4	YAP1	1.22E-02	4.20E-02	NA
38	RME1	YAP1	5.01E-03	4.20E-02	NA
39	SWI4	YAP1	6.01E-03	4.20E-02	NA

Supplementary Table A8 $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 6$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	PDR1	RAP1	3.41E-09	1.19E-08	NA
2	FKH2	NDD1	3.26E-07	5.36E-08	Kumar *et al.*, 2000
3	FHL1	PDR1	9.50E-07	1.00E-06	NA
4	RAP1	YAP5	1.20E-10	2.33E-05	NA
5	ABF1	REB1	1.67E-04	5.97E-05	Schuller *et al.*, 1994
6	FKH2	SWI5	1.43E-04	3.24E-04	Pic *et al.*, 2000
7	HIR1	HIR2	8.40E-06	7.07E-04	Loy *et al.*, 1999
8	PHO4	SKN7	2.07E-03	3.10E-04	NA
9	MBP1	PHO4	2.33E-03	2.07E-03	NA
10	ACE2	PHO4	4.59E-03	2.07E-03	NA
11	ACE2	FKH2	6.49E-03	3.62E-04	Pic *et al.*, 2000
12	MBP1	SWI6	7.34E-03	4.01E-07	Koch *et al.*, 1993
13	FKH2	RME1	6.04E-03	7.58E-03	NA
14	RME1	SKN7	7.58E-03	5.26E-03	NA
15	RME1	SWI4	7.58E-03	7.81E-03	NA
16	ACE2	YAP1	1.02E-03	8.12E-03	NA
17	MBP1	YAP1	6.38E-04	8.12E-03	NA
18	SKN7	YAP1	8.70E-05	8.12E-03	Lee *et al.*, 1999
19	SWI5	YAP1	2.96E-04	8.12E-03	NA
20	FKH2	YAP1	2.09E-04	1.01E-02	NA
21	RME1	SWI5	7.58E-03	1.01E-02	NA
22	ACE2	SWI5	1.26E-02	9.18E-04	Zhu *et al.*, 2000
23	ACE2	NDD1	1.36E-02	4.88E-04	NA
24	FKH2	PHO4	6.04E-03	1.44E-02	NA
25	NDD1	PHO4	3.81E-03	1.44E-02	NA
26	PHO4	RME1	1.44E-02	7.58E-03	NA
27	PHO4	SWI4	1.44E-02	7.81E-03	NA
28	PHO4	SWI5	1.44E-02	1.01E-02	NA
29	ARG80	ARG81	1.01E-02	1.87E-02	Mamnun *et al.*, 2002
30	ACE2	RME1	2.07E-02	7.58E-03	NA
31	RAP1	SFP1	2.70E-02	1.01E-02	NA
32	MBP1	MSN4	3.53E-02	2.02E-02	NA

Supplementary Table A9 $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 7$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	MBP1	SWI6	3.29E-11	2.34E-17	Koch *et al.*, 1993
2	FKH2	NDD1	1.04E-08	3.75E-12	Kumar *et al.*, 2000
3	PDR1	RAP1	3.41E-09	3.92E-08	NA
4	FHL1	PDR1	2.16E-06	1.00E-06	NA
5	RAP1	YAP5	2.12E-08	2.33E-05	NA
6	ACE2	FKH2	1.86E-04	1.05E-05	Pic *et al.*, 2000
7	ABF1	REB1	2.13E-04	6.84E-05	Schuller *et al.*, 1994
8	FKH2	SWI5	2.36E-04	3.24E-04	Pic *et al.*, 2000
9	PHO4	SKN7	3.09E-03	5.09E-04	NA
10	MBP1	PHO4	3.53E-03	3.09E-03	NA
11	NDD1	SWI4	3.77E-04	4.95E-03	NA
12	ACE2	PHO4	6.86E-03	3.09E-03	NA
13	RME1	SKN7	7.58E-03	6.45E-03	NA
14	FKH2	RME1	8.00E-03	7.58E-03	NA
15	FKH1	NDD1	8.42E-03	1.52E-03	Kumar *et al.*, 2000
16	RME1	SWI4	7.58E-03	8.42E-03	NA
17	FKH1	FKH2	9.11E-03	9.03E-03	Koranda *et al.*, 2000; Spector *et al.*, 1997
18	RME1	SWI5	7.58E-03	1.01E-02	NA
19	HIR1	HIR2	8.40E-06	1.42E-02	Loy *et al.*, 1999
20	ACE2	YAP1	1.42E-03	1.52E-02	NA
21	MBP1	YAP1	8.94E-04	1.52E-02	NA
22	SKN7	YAP1	1.20E-04	1.52E-02	Lee *et al.*, 1999
23	SWI5	YAP1	2.96E-04	1.52E-02	NA
24	FKH2	PHO4	8.00E-03	1.73E-02	NA
25	NDD1	PHO4	4.46E-03	1.73E-02	NA
26	PHO4	RME1	1.73E-02	7.58E-03	NA
27	PHO4	SWI4	1.73E-02	8.42E-03	NA
28	PHO4	SWI5	1.73E-02	1.01E-02	NA
29	ARG80	ARG81	1.01E-02	1.87E-02	Mamnun *et al.*, 2002
30	FKH2	YAP1	3.66E-04	1.91E-02	NA
31	ACE2	RME1	2.50E-02	7.58E-03	NA
32	ACE2	NDD1	2.77E-02	2.33E-03	NA
33	NDD1	SKN7	3.55E-02	2.13E-02	NA
34	RAP1	SFP1	3.67E-02	1.01E-02	NA

Supplementary Table A10 $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 10$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	FKH2	NDD1	2.34E-07	3.75E-12	Kumar *et al.*, 2000
2	PDR1	RAP1	3.41E-08	3.44E-07	NA
3	MBP1	SWI6	3.22E-11	4.78E-06	Koch *et al.*, 1993
4	FHL1	PDR1	5.52E-06	4.75E-05	NA
5	ABF1	REB1	4.33E-04	8.90E-05	Schuller *et al.*, 1994
6	FKH2	SWI5	2.73E-04	7.16E-04	Pic *et al.*, 2000
7	ACE2	FKH2	1.82E-03	1.78E-05	Pic *et al.*, 2000
8	HIR1	HIR2	5.95E-06	5.61E-03	Loy *et al.*, 1999
9	ACE2	SWI5	1.10E-02	5.86E-05	Zhu *et al.*, 2000
10	FKH2	RME1	8.69E-03	1.40E-02	NA
11	RME1	SWI4	1.40E-02	1.09E-02	NA
12	ACE2	YAP1	3.03E-03	1.52E-02	NA
13	MBP1	YAP1	1.30E-03	1.52E-02	NA
14	SKN7	YAP1	5.76E-04	1.52E-02	Lee *et al.*, 1999
15	SWI5	YAP1	6.45E-04	1.52E-02	NA
16	RME1	SWI5	1.40E-02	1.59E-02	NA
17	RME1	SKN7	1.40E-02	1.68E-02	NA
18	FKH2	YAP1	4.31E-04	1.91E-02	NA
19	ACE2	PHO4	1.69E-02	2.34E-02	NA
20	MBP1	PHO4	5.49E-03	2.34E-02	NA
21	PHO4	SKN7	2.34E-02	4.97E-03	NA
22	ACE2	RME1	3.85E-02	1.40E-02	NA
23	FKH2	PHO4	8.69E-03	4.52E-02	NA
24	NDD1	PHO4	4.46E-03	4.52E-02	NA
25	PHO4	RME1	4.52E-02	1.40E-02	NA
26	PHO4	SWI4	4.52E-02	1.09E-02	NA
27	PHO4	SWI5	4.52E-02	1.59E-02	NA

Supplementary Table A11 $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 15$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	PDR1	RAP1	1.18E-07	1.50E-06	NA
2	FHL1	PDR1	2.14E-05	5.62E-04	NA
3	ABF1	REB1	1.35E-03	1.30E-04	Schuller *et al.*, 1994
4	FKH2	SWI5	8.38E-04	8.51E-03	Pic *et al.*, 2000
5	MBP1	SWI6	1.08E-17	1.28E-02	Koch *et al.*, 1993
6	ACE2	SWI5	1.34E-02	2.34E-03	Zhu *et al.*, 2000
7	FKH2	RME1	1.63E-02	3.54E-02	NA
8	RME1	SWI4	3.54E-02	2.16E-02	NA
9	FKH2	SKN7	1.63E-02	3.73E-02	NA

Supplementary Table A12 $P_B < 0.0001$, $O_{\min} = 3$, $I_{\min} = 1$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	ACE2	SKN7	5.08E-07	3.37E-07	NA
2	NDD1	SWI4	1.70E-05	1.84E-07	NA
3	MBP1	NDD1	2.01E-06	1.86E-05	NA
4	MBP1	PHO4	6.74E-05	2.65E-05	NA
5	MBP1	SKN7	5.03E-06	1.09E-04	NA
6	ACE2	PHO4	1.80E-04	2.65E-05	NA
7	PHO4	SKN7	2.65E-05	2.33E-04	NA
8	NDD1	SWI5	5.46E-04	4.25E-04	NA
9	SKN7	SWI5	1.81E-03	1.44E-04	NA
10	SWI4	SWI5	1.99E-03	1.96E-03	Cosma *et al.*, 1999
11	NDD1	PHO4	2.18E-03	2.17E-03	NA
12	FKH2	MCM1	1.60E-03	2.29E-03	Spector *et al.*, 1997
13	MCM1	NDD1	2.29E-03	1.05E-07	Kumar *et al.*, 2000
14	PHO4	SWI5	2.17E-03	2.35E-03	NA
15	PHO4	SWI4	2.17E-03	2.55E-03	NA
16	SWI4	SWI6	1.87E-04	5.60E-03	Koch *et al.*, 1993
17	CIN5	PHD1	6.42E-03	2.98E-03	NA
18	MCM1	SWI4	7.52E-03	1.13E-04	NA
19	HIR1	HIR2	8.06E-04	1.01E-02	Loy *et al.*, 1999
20	ACE2	NDD1	1.07E-02	3.96E-04	NA
21	PDR1	YAP5	1.52E-03	1.29E-02	NA
22	NRG1	PHD1	1.21E-04	1.86E-02	NA
23	PHD1	SKN7	3.21E-03	1.92E-02	NA
24	MBP1	SWI5	1.63E-02	3.08E-02	NA
25	ACE2	MBP1	3.83E-02	5.24E-05	NA

Supplementary Table A13 $P_B < 0.0001$, $O_{\min} = 3$, $I_{\min} = 2$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	SWI4	SWI6	7.18E-10	7.79E-07	Koch *et al.*, 1993
2	ACE2	SKN7	1.36E-09	1.46E-06	NA
3	ACE2	SWI5	1.47E-05	4.06E-04	Zhu *et al.*, 2000
4	FKH2	MCM1	4.94E-05	7.74E-04	Spector *et al.*, 1997
5	MCM1	NDD1	7.74E-04	4.02E-09	Kumar *et al.*, 2000
6	MBP1	SWI4	1.08E-03	2.85E-08	NA
7	ACE2	PHO4	2.07E-03	5.79E-03	NA
8	PHO4	SKN7	5.79E-03	6.52E-03	NA
9	MBP1	PHO4	1.19E-02	5.79E-03	NA
10	ACE2	NDD1	1.59E-02	5.85E-03	NA
11	HIR1	HIR2	1.90E-02	1.01E-02	Loy *et al.*, 1999
12	MBP1	SKN7	2.13E-02	4.90E-03	Bouquin *et al.*, 1999
13	ACE2	SWI4	4.21E-02	2.29E-02	NA
14	MCM1	SWI4	4.50E-02	8.75E-03	NA
15	PHD1	SKN7	9.31E-03	4.95E-02	NA

Supplementary Table A14 $P_B < 0.0001$, $O_{\min} = 3$, $I_{\min} = 3$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	ACE2	SKN7	1.06E-03	1.42E-04	NA
2	SWI4	SWI6	2.67E-02	4.37E-04	Koch *et al.*, 1993

Supplementary Table A15 $P_B < 0.0001$, $O_{\min} = 3$, $I_{\min} = 5$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	ACE2	PHO4	2.48E-04	5.36E-04	NA
2	PHO4	SKN7	5.36E-04	1.18E-03	NA
3	FHL1	YAP5	2.94E-03	3.78E-03	NA
4	MBP1	PHO4	6.15E-03	5.36E-04	NA
5	NDD1	PHO4	2.14E-03	6.36E-03	NA
6	PHO4	SWI4	6.36E-03	4.63E-03	NA
7	PHO4	SWI5	6.36E-03	4.48E-03	NA
8	ACE2	SWI5	5.09E-03	7.03E-03	Zhu *et al.*, 2000
9	ACE2	SKN7	8.82E-07	8.82E-03	NA

Supplementary Table A16 $P_B < 0.0001$, $O_{\min} = 3$, $I_{\min} = 10$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	ACE2	PHO4	9.96E-04	3.65E-03	NA
2	FHL1	YAP5	3.21E-03	3.78E-03	NA
3	PHO4	SKN7	3.65E-03	1.05E-02	NA
4	NDD1	PHO4	2.62E-03	1.37E-02	NA
5	PHO4	SWI4	1.37E-02	7.96E-03	NA
6	PHO4	SWI5	1.37E-02	8.06E-03	NA
7	MBP1	PHO4	4.59E-02	3.65E-03	NA

Supplementary Table A17 D_G, $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 3$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	HIR1	HIR2	4.80E-11	5.94E-08	Loy *et al.*, 1999
2	RAP1	YAP5	5.70E-04	6.45E-10	NA
3	MCM1	SWI4	1.90E-03	1.93E-03	NA
4	FHL1	PDR1	4.29E-03	1.02E-05	NA
5	RAP1	SFP1	5.09E-03	2.75E-03	NA
6	PHO4	SKN7	6.16E-03	1.17E-03	NA
7	NDD1	SKN7	8.07E-03	1.78E-03	NA
8	NDD1	SWI5	1.74E-03	9.41E-03	NA
9	ACE2	NDD1	1.04E-02	4.04E-06	NA
10	FKH1	NDD1	1.76E-02	6.91E-03	Kumar *et al.*, 2000
11	ABF1	REB1	2.00E-02	1.04E-02	Schuller *et al.*, 1994
12	NRG1	YAP6	1.55E-04	2.01E-02	NA
13	HSF1	RAP1	2.38E-02	1.71E-02	NA
14	PDR1	RAP1	2.44E-02	1.86E-02	NA
15	MBP1	PHO4	3.02E-02	6.16E-03	NA
16	FKH1	FKH2	1.76E-02	3.53E-02	Koranda *et al.*, 2000; Spector *et al.*, 1997
17	NRG1	ROX1	5.37E-06	3.78E-02	NA
18	SWI5	YAP1	6.37E-03	4.60E-02	NA
19	NRG1	SWI5	1.52E-02	4.69E-02	NA

Supplementary Table A18 D_G, $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 5$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	PDR1	RAP1	2.08E-09	7.93E-11	NA
2	FHL1	RAP1	7.70E-09	1.16E-33	NA
3	HIR1	HIR2	6.99E-11	1.19E-06	Loy *et al.*, 1999
4	FHL1	PDR1	6.25E-06	7.51E-07	NA
5	RAP1	YAP5	3.67E-06	2.65E-05	NA
6	PHO4	SKN7	1.33E-04	1.02E-04	NA
7	ACE2	NDD1	6.29E-04	2.54E-06	NA
8	MBP1	PHO4	1.11E-03	1.33E-04	NA
9	ABF1	REB1	2.19E-03	8.39E-04	Schuller *et al.*, 1994
10	ACE2	PHO4	2.24E-03	1.33E-04	NA
11	NDD1	SWI5	1.37E-03	3.40E-03	NA
12	RME1	SKN7	5.22E-05	4.19E-03	NA
13	NDD1	PHO4	4.35E-03	5.89E-03	NA
14	PHO4	RME1	5.89E-03	5.22E-05	NA
15	PHO4	SWI5	5.89E-03	6.91E-03	NA
16	RME1	SWI5	5.22E-05	6.91E-03	NA
17	ACE2	MBP1	7.80E-03	7.48E-03	NA
18	FKH2	PHO4	8.30E-03	5.89E-03	NA
19	FKH2	RME1	8.30E-03	5.22E-05	NA
20	NDD1	SWI4	3.02E-04	8.70E-03	NA
21	RAP1	SFP1	1.12E-02	4.34E-03	NA
22	PHO4	SWI4	5.89E-03	1.16E-02	NA
23	RME1	SWI4	5.22E-05	1.16E-02	NA
24	ACE2	YAP1	1.77E-03	1.16E-02	NA
25	SWI5	YAP1	1.50E-04	1.16E-02	NA
26	MBP1	SWI5	1.64E-02	3.40E-03	NA
27	NRG1	ROX1	4.37E-04	1.74E-02	NA
28	FKH2	SWI5	1.97E-02	1.45E-02	Pic *et al.*, 2000
29	ACE2	RME1	2.30E-02	5.22E-05	NA
30	ACE2	FKH2	2.49E-02	2.23E-03	Pic *et al.*, 2000
31	FKH2	NDD1	4.26E-02	2.56E-03	Kumar *et al.*, 2000
32	FKH2	YAP1	3.03E-03	4.64E-02	NA

Supplementary Table A19 D_G, $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 10$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	PDR1	RAP1	1.04E-08	1.29E-08	NA
2	FHL1	RAP1	3.12E-16	1.72E-07	NA
3	FHL1	PDR1	1.14E-04	7.58E-06	NA
4	HIR1	HIR2	5.76E-08	2.93E-04	Loy *et al.*, 1999
5	RAP1	YAP5	3.24E-04	1.21E-04	NA
6	MBP1	SWI6	2.40E-06	1.15E-03	Koch *et al.*, 1993
7	FKH2	SWI5	1.70E-03	2.48E-03	Pic *et al.*, 2000
8	MBP1	PHO4	4.24E-03	3.23E-03	NA
9	PHO4	SKN7	3.23E-03	6.71E-03	NA
10	FKH2	NDD1	6.76E-03	1.68E-05	Kumar *et al.*, 2000
11	ACE2	FKH2	1.32E-02	9.53E-05	Pic *et al.*, 2000
12	ABF1	REB1	1.68E-02	3.52E-03	Schuller *et al.*, 1994
13	FKH2	RME1	2.04E-02	2.66E-04	NA
14	FKH2	PHO4	2.04E-02	2.72E-02	NA
15	NDD1	PHO4	8.63E-03	2.72E-02	NA
16	PHO4	RME1	2.72E-02	2.66E-04	NA
17	ACE2	PHO4	2.91E-02	3.23E-03	NA
18	RME1	SWI5	2.66E-04	2.99E-02	NA
19	RME1	SWI4	2.66E-04	3.08E-02	NA
20	ACE2	YAP1	2.22E-02	3.51E-02	NA
21	SWI5	YAP1	2.48E-03	3.51E-02	NA
22	NDD1	SWI4	6.87E-04	3.72E-02	NA
23	RME1	SKN7	2.66E-04	3.83E-02	NA
24	CIN5	NRG1	4.77E-02	1.53E-02	NA

Supplementary Table A20 D_{CD}, $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 1$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	HIR1	HIR2	2.46E-17	1.02E-16	Loy *et al.*, 1999
2	PDR1	RAP1	3.01E-20	4.46E-11	NA
3	PHO4	SKN7	2.21E-08	5.24E-12	NA
4	ACE2	FKH2	1.03E-06	1.03E-07	Pic *et al.*, 2000
5	ACE2	PHO4	2.95E-06	2.21E-08	NA
6	MBP1	PHO4	4.36E-06	2.21E-08	NA
7	FHL1	PDR1	6.50E-06	3.02E-29	NA
8	RME1	SWI5	1.29E-06	2.22E-05	NA
9	RAP1	YAP5	3.70E-05	2.97E-33	NA
10	RME1	SWI4	1.29E-06	4.14E-04	NA
11	ACE2	YAP1	2.60E-06	4.90E-04	NA
12	SWI5	YAP1	8.10E-12	4.90E-04	NA
13	NDD1	PHO4	9.71E-06	5.72E-04	NA
14	PHO4	RME1	5.72E-04	1.29E-06	NA
15	PHO4	SWI4	5.72E-04	4.14E-04	NA
16	PHO4	SWI5	5.72E-04	2.22E-05	NA
17	FKH2	SWI5	7.05E-04	3.71E-07	Pic *et al.*, 2000
18	AZF1	CIN5	7.83E-04	1.10E-04	NA
19	ABF1	REB1	8.35E-04	1.01E-04	Schuller *et al.*, 1994
20	NDD1	SWI4	1.37E-08	1.01E-03	NA
21	FKH2	PHO4	1.67E-03	5.72E-04	NA
22	FKH2	RME1	1.67E-03	1.29E-06	NA
23	RME1	SKN7	7.31E-04	1.95E-03	NA
24	NDD1	RME1	9.62E-04	2.55E-03	NA
25	ACE2	RME1	3.42E-03	1.29E-06	NA
26	FHL1	RAP1	3.54E-03	1.43E-85	NA
27	MBP1	MSN4	4.27E-03	3.55E-05	NA
28	MBP1	SWI6	4.03E-03	7.45E-03	Koch *et al.*, 1993
29	MBP1	YAP1	4.29E-04	9.10E-03	NA
30	ABF1	INO4	1.21E-02	9.07E-03	NA
31	RAP1	SFP1	1.58E-02	1.07E-02	NA
32	FHL1	GAT3	1.59E-02	3.86E-14	NA
33	ABF1	PHO4	2.09E-02	8.25E-04	NA
34	PHO4	YAP1	5.72E-04	2.95E-02	NA
35	RME1	YAP1	1.29E-06	2.95E-02	NA
36	SWI4	YAP1	4.14E-04	2.95E-02	NA
37	MBP1	SKN7	3.58E-02	1.73E-06	Bouquin *et al.*, 1999
38	HAP3	HAP4	8.26E-04	3.84E-02	Riego *et al.*, 2002
39	NRG1	YAP6	6.05E-06	3.88E-02	NA
40	MAC1	ROX1	4.07E-02	7.37E-04	NA

Supplementary Table A21 D_{CD}, $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 2$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	HIR1	HIR2	1.65E-12	9.30E-13	Loy *et al.*, 1999
2	PDR1	RAP1	2.15E-08	2.29E-07	NA
3	FHL1	RAP1	3.00E-06	6.16E-48	NA
4	RAP1	YAP5	6.72E-05	1.14E-10	NA
5	ACE2	FKH2	9.46E-05	2.70E-05	Pic *et al.*, 2000
6	PHO4	SKN7	9.90E-05	1.21E-08	NA
7	FHL1	PDR1	1.54E-04	4.31E-12	NA
8	ACE2	PHO4	1.81E-04	9.90E-05	NA
9	MBP1	PHO4	6.55E-04	9.90E-05	NA
10	ACE2	YAP1	1.61E-04	1.22E-03	NA
11	SWI5	YAP1	8.23E-06	1.22E-03	NA
12	MCM1	SWI4	2.01E-04	1.78E-03	NA
13	ABF1	REB1	4.33E-03	9.22E-04	Schuller *et al.*, 1994
14	RME1	SWI5	2.29E-05	6.01E-03	NA
15	ARO80	GAT3	6.54E-03	5.77E-04	NA
16	ARO80	YAP5	6.54E-03	5.63E-03	NA
17	ARO80	GAL4	6.54E-03	7.78E-03	NA
18	GAL4	GAT3	8.04E-03	1.79E-04	NA
19	ACE2	MSN4	7.19E-03	8.79E-03	NA
20	MSN4	NDD1	8.79E-03	1.32E-03	NA
21	RME1	SWI4	2.29E-05	9.31E-03	NA
22	FKH2	RME1	1.30E-02	2.29E-05	NA
23	RME1	SKN7	3.50E-03	1.35E-02	NA
24	FKH2	SWI5	1.66E-02	3.37E-03	Pic *et al.*, 2000
25	FKH2	PHO4	1.30E-02	1.72E-02	NA
26	NDD1	PHO4	1.86E-03	1.72E-02	NA
27	PHO4	RME1	1.72E-02	2.29E-05	NA
28	PHO4	SWI4	1.72E-02	9.31E-03	NA
29	PHO4	SWI5	1.72E-02	6.01E-03	NA
30	MBP1	YAP1	1.43E-02	1.73E-02	NA
31	NRG1	YAP6	8.32E-04	1.81E-02	NA
32	ACE2	RME1	1.95E-02	2.29E-05	NA
33	MBP1	SWI6	2.19E-02	5.78E-03	Koch *et al.*, 1993
34	ABF1	INO4	1.88E-02	2.40E-02	NA
35	ACE2	NDD1	2.53E-02	4.46E-05	NA
36	HSF1	YAP1	2.98E-02	2.31E-02	NA
37	GAL4	HAP4	7.78E-03	3.04E-02	NA
38	FHL1	GAT3	3.39E-02	8.68E-07	NA
39	ABF1	PHO4	3.39E-02	7.42E-03	NA
40	MBP1	MSN4	3.76E-02	2.08E-02	NA
41	NDD1	RME1	4.39E-02	1.17E-02	NA
42	PHO4	YAP1	1.72E-02	4.39E-02	NA
43	RME1	YAP1	2.29E-05	4.39E-02	NA
44	SWI4	YAP1	9.31E-03	4.39E-02	NA

Supplementary Table A22 D_{CD}, $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 3$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	HIR1	HIR2	1.51E-10	5.82E-08	Loy *et al.*, 1999
2	RAP1	YAP5	2.01E-06	5.28E-10	NA
3	PDR1	RAP1	1.47E-04	3.33E-06	NA
4	PHO4	SKN7	3.96E-04	4.62E-06	NA
5	FHL1	PDR1	9.86E-04	1.50E-06	NA
6	FHL1	RAP1	1.13E-03	8.51E-37	NA
7	NRG1	YAP6	6.66E-04	2.31E-03	NA
8	RME1	SKN7	7.15E-04	3.57E-03	NA
9	MCM1	SWI4	4.36E-03	5.76E-03	NA
10	RAP1	SFP1	6.16E-03	2.67E-03	NA
11	MBP1	PHO4	6.22E-03	3.96E-04	NA
12	MCM1	YJL206C	7.28E-03	6.56E-03	NA
13	NDD1	SWI5	8.22E-04	9.57E-03	NA
14	ARG81	CIN5	4.31E-03	9.78E-03	NA
15	ABF1	REB1	1.05E-02	3.41E-03	NA
16	ACE2	NDD1	1.05E-02	5.21E-08	NA
17	ACE2	PHO4	1.14E-02	3.96E-04	NA
18	GAL4	GAT3	1.17E-02	1.37E-03	NA
19	MSN4	NDD1	1.26E-02	3.87E-03	NA
20	ACE2	YAP1	1.14E-02	1.35E-02	NA
21	SWI5	YAP1	3.83E-04	1.35E-02	NA
22	ARG80	GCN4	2.34E-03	1.71E-02	NA
23	NRG1	ROX1	9.14E-09	1.79E-02	NA
24	ACE2	MSN4	2.16E-02	1.26E-02	NA
25	ARO80	GAL4	2.26E-02	1.17E-02	NA
26	ARO80	GAT3	2.26E-02	1.37E-03	NA
27	ARO80	YAP5	2.26E-02	1.40E-02	NA
28	ABF1	INO4	2.37E-02	2.40E-02	NA
29	MBP1	SMP1	2.51E-02	3.42E-03	NA
30	RME1	SWI5	7.15E-04	2.95E-02	NA
31	BAS1	HAP2	1.69E-03	2.97E-02	NA
32	NDD1	PHO4	9.36E-03	3.05E-02	NA
33	PHO4	RME1	3.05E-02	7.15E-04	NA
34	FKH2	RME1	3.10E-02	7.15E-04	NA
35	RME1	SWI4	7.15E-04	3.35E-02	NA
36	FHL1	YAP5	3.51E-02	6.35E-13	NA
37	NRG1	SWI5	3.02E-03	3.56E-02	NA
38	FZF1	SWI5	6.96E-03	4.03E-02	NA
39	GAL4	HAP4	1.17E-02	4.10E-02	NA
40	ABF1	PHO4	4.38E-02	1.07E-02	NA
41	CAD1	CIN5	1.77E-02	4.55E-02	NA

Supplementary Table A23 D_{CD}, $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 5$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	PDR1	RAP1	3.95E-09	4.25E-11	NA
2	FHL1	RAP1	4.61E-07	1.92E-25	NA
3	RAP1	YAP5	4.27E-07	9.96E-06	NA
4	PHO4	SKN7	2.12E-05	4.35E-07	NA
5	HIR1	HIR2	1.00E-08	2.70E-05	Loy *et al.*, 1999
6	FHL1	PDR1	3.86E-05	1.38E-06	NA
7	ACE2	PHO4	8.82E-05	2.12E-05	NA
8	ABF1	REB1	1.26E-04	2.48E-05	Schuller *et al.*, 1994
9	MBP1	PHO4	1.35E-04	2.12E-05	NA
10	RME1	SKN7	8.63E-07	2.45E-04	NA
11	ACE2	NDD1	5.87E-04	4.92E-07	NA
12	FKH2	RME1	1.12E-03	8.63E-07	NA
13	RME1	SWI5	8.63E-07	1.36E-03	NA
14	FKH2	PHO4	1.12E-03	2.12E-03	NA
15	NDD1	PHO4	5.97E-04	2.12E-03	NA
16	PHO4	RME1	2.12E-03	8.63E-07	NA
17	PHO4	SWI5	2.12E-03	1.36E-03	NA
18	PHO4	SWI4	2.12E-03	2.77E-03	NA
19	RME1	SWI4	8.63E-07	2.77E-03	NA
20	ACE2	YAP1	1.01E-04	2.91E-03	NA
21	SWI5	YAP1	8.76E-06	2.91E-03	NA
22	NDD1	SWI5	8.78E-04	3.48E-03	NA
23	MBP1	SWI6	3.97E-03	3.39E-04	Koch *et al.*, 1993
24	ACE2	RME1	4.20E-03	8.63E-07	NA
25	ACE2	MBP1	6.04E-03	8.63E-03	NA
26	RLM1	SMP1	3.51E-03	1.06E-02	NA
27	NRG1	ROX1	3.99E-05	1.09E-02	NA
28	NDD1	SWI4	4.33E-04	1.48E-02	NA
29	MCM1	YJL206C	1.58E-02	2.34E-03	NA
30	MSN4	NDD1	2.01E-02	1.30E-02	NA
31	RAP1	SFP1	2.13E-02	5.39E-03	NA
32	ARG81	CIN5	4.31E-03	2.22E-02	NA
33	FZF1	NRG1	1.63E-02	2.42E-02	NA
34	PHO4	YAP1	2.12E-03	2.51E-02	NA
35	RME1	YAP1	8.63E-07	2.51E-02	NA
36	SWI4	YAP1	2.77E-03	2.51E-02	NA
37	FKH2	SWI5	2.60E-02	2.03E-02	Pic *et al.*, 2000
38	MBP1	SWI5	2.63E-02	3.48E-03	NA
39	ACE2	MSN4	2.77E-02	2.01E-02	NA
40	BAS1	HAP2	2.36E-02	2.97E-02	NA
41	ARG80	GCN4	2.34E-03	3.02E-02	NA
42	FHL1	YAP5	3.56E-02	5.73E-05	NA
43	FKH2	YAP1	5.23E-04	3.66E-02	NA
44	MBP1	SMP1	3.98E-02	1.06E-02	NA

Supplementary Table A24 D_{CD}, $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 10$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	PDR1	RAP1	2.39E-08	1.79E-07	NA
2	FHL1	RAP1	1.42E-12	2.40E-06	NA
3	FHL1	PDR1	4.44E-04	1.15E-05	NA
4	RAP1	YAP5	2.94E-04	7.21E-04	NA
5	MBP1	PHO4	7.98E-04	1.29E-03	NA
6	PHO4	SKN7	1.29E-03	8.23E-04	NA
7	MBP1	SWI6	1.01E-07	1.67E-03	Koch *et al.*, 1993
8	ABF1	REB1	1.93E-03	1.64E-04	NA
9	ACE2	PHO4	2.53E-03	1.29E-03	NA
10	FKH2	RME1	4.32E-03	3.03E-04	NA
11	FKH2	SWI5	1.97E-03	5.14E-03	Pic *et al.*, 2000
12	HIR1	HIR2	5.44E-07	5.56E-03	Loy *et al.*, 1999
13	RME1	SWI4	3.03E-04	1.02E-02	NA
14	CIN5	GCN4	6.85E-03	1.09E-02	NA
15	RME1	SWI5	3.03E-04	1.10E-02	NA
16	RME1	SKN7	3.03E-04	1.20E-02	NA
17	ACE2	YAP1	3.01E-03	1.21E-02	NA
18	SWI5	YAP1	5.85E-04	1.21E-02	NA
19	FKH2	PHO4	4.32E-03	1.36E-02	NA
20	NDD1	PHO4	2.33E-03	1.36E-02	NA
21	PHO4	RME1	1.36E-02	3.03E-04	NA
22	PHO4	SWI4	1.36E-02	1.02E-02	NA
23	PHO4	SWI5	1.36E-02	1.10E-02	NA
24	NDD1	SWI4	1.93E-04	1.77E-02	NA
25	ACE2	FKH2	2.06E-02	1.72E-04	Pic *et al.*, 2000
26	ACE2	RME1	2.31E-02	3.03E-04	NA
27	MSN4	NDD1	2.40E-02	2.03E-02	NA
28	FKH2	NDD1	3.39E-02	1.18E-04	Kumar *et al.*, 2000
29	ACE2	MSN4	4.59E-02	2.40E-02	NA

Supplementary Table A25 $D(i,j,3)$, $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 1$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	ACE2	FKH2	1.39E-12	3.81E-10	Pic *et al.*, 2000
2	HIR1	HIR2	3.98E-08	2.31E-09	Loy *et al.*, 1999
3	MBP1	SKN7	7.03E-08	3.59E-10	Bouquin *et al.*, 1999
4	NDD1	SWI4	3.52E-12	1.49E-07	NA
5	NDD1	RME1	1.69E-07	1.70E-06	NA
6	SWI4	SWI5	1.95E-06	3.84E-09	Cosma *et al.*, 1999
7	NRG1	PHD1	5.27E-12	2.11E-06	NA
8	FHL1	PDR1	2.54E-06	5.44E-25	NA
9	FKH2	SWI4	5.45E-06	1.41E-17	NA
10	PHD1	SKN7	6.20E-10	1.13E-05	NA
11	MBP1	RME1	8.11E-05	1.70E-06	NA
12	PDR1	RAP1	2.64E-15	9.81E-05	NA
13	FKH2	SWI5	1.01E-04	5.51E-07	Pic *et al.*, 2000
14	MBP1	MSN4	2.35E-04	3.60E-05	NA
15	FHL1	GAT3	2.47E-04	5.00E-11	NA
16	ACE2	PHO4	7.06E-05	2.63E-04	NA
17	MBP1	PHO4	2.14E-04	2.63E-04	NA
18	PHO4	SKN7	2.63E-04	2.64E-05	NA
19	NDD1	YAP1	8.71E-09	3.11E-04	NA
20	ACE2	YAP1	5.46E-05	5.13E-04	NA
21	SWI5	YAP1	6.84E-06	5.13E-04	NA
22	MBP1	YAP1	1.46E-04	5.48E-04	NA
23	HAP3	HAP5	6.06E-04	5.16E-05	McNabb *et al.*, 1995
24	HAP2	HAP5	9.58E-04	5.16E-05	McNabb *et al.*, 1995
25	ACE2	SWI5	1.86E-03	3.28E-27	Zhu *et al.*, 2000
26	FHL1	RAP1	2.86E-03	8.47E-198	NA
27	RME1	SKN7	4.11E-03	1.01E-03	NA
28	ACE2	RME1	3.37E-03	4.59E-03	NA
29	RME1	SWI4	4.59E-03	1.55E-03	NA
30	RME1	SWI5	4.59E-03	8.17E-04	NA
31	MAC1	ROX1	5.46E-03	1.04E-03	NA
32	FKH2	RME1	6.84E-03	4.59E-03	NA
33	SKN7	SWI4	7.81E-03	2.41E-06	NA
34	CIN5	GCN4	3.17E-04	8.42E-03	NA
35	FKH2	PHO4	6.84E-03	8.62E-03	NA
36	NDD1	PHO4	7.03E-04	8.62E-03	NA
37	PHO4	RME1	8.62E-03	4.59E-03	NA
38	PHO4	SWI4	8.62E-03	1.55E-03	NA
39	PHO4	SWI5	8.62E-03	8.17E-04	NA
40	HAP2	HAP3	9.33E-03	1.81E-03	McNabb *et al.*, 1995
41	ABF1	REB1	1.08E-02	8.95E-03	Schuller *et al.*, 1994
42	SKN7	YAP6	1.23E-02	5.59E-03	NA
43	GAL4	YAP5	1.35E-02	5.37E-05	NA
44	FKH2	YAP1	1.18E-03	1.35E-02	NA
45	PHO4	YAP1	8.62E-03	1.58E-02	NA
46	RME1	YAP1	4.59E-03	1.58E-02	NA
47	SWI4	YAP1	1.55E-03	1.58E-02	NA
48	SKN7	SWI5	1.75E-02	4.74E-05	NA
49	STE12	SWI6	1.40E-02	1.95E-02	NA
50	GCR1	GCR2	2.48E-02	4.35E-03	Turkel and Bisson, 1999; Uemura *et al.*, 1997
51	RLM1	ROX1	1.93E-06	3.32E-02	NA
52	FKH2	SKN7	4.27E-02	3.49E-03	NA
53	SWI5	SWI6	2.90E-03	4.29E-02	NA
54	CBF1	MET4	2.08E-02	4.71E-02	Wheeler *et al.*, 2003
55	MBP1	NDD1	4.75E-02	5.46E-08	NA

Supplementary Table A26 $D(i,j,3)$, $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 2$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	FHL1	RAP1	3.86E-13	3.47E-91	NA
2	HIR1	HIR2	1.45E-06	1.51E-08	Loy *et al.*, 1999
3	ACE2	FKH2	2.90E-08	1.91E-06	Pic *et al.*, 2000
4	MCM1	SWI4	1.28E-06	9.78E-06	NA
5	NRG1	PHD1	6.96E-05	2.02E-05	NA
6	MBP1	SKN7	7.96E-05	1.64E-11	Bouquin *et al.*, 1999
7	FHL1	PDR1	1.65E-04	9.68E-16	NA
8	FKH2	MBP1	5.15E-06	2.75E-04	NA
9	NDD1	RME1	3.11E-04	5.06E-04	NA
10	HAP2	HAP5	2.22E-04	5.69E-04	McNabb *et al.*, 1995
11	HAP3	HAP5	2.29E-04	5.69E-04	McNabb *et al.*, 1995
12	FKH2	MCM1	7.09E-04	5.40E-11	Spector *et al.*, 1997
13	NDD1	YAP1	2.72E-04	8.54E-04	NA
14	MBP1	MSN4	1.19E-03	3.34E-04	NA
15	FHL1	GAT3	1.23E-03	1.30E-09	NA
16	ACE2	YAP1	1.28E-03	9.12E-04	NA
17	SWI5	YAP1	1.81E-03	9.12E-04	NA
18	MBP1	MCM1	1.91E-03	2.34E-06	NA
19	NRG1	YAP6	5.45E-07	2.96E-03	NA
20	MAC1	ROX1	3.56E-03	1.33E-06	NA
21	MCM1	SWI6	1.52E-04	4.41E-03	NA
22	MBP1	YAP1	6.54E-03	9.98E-04	NA
23	NDD1	SKN7	1.05E-02	9.10E-03	NA
24	FKH2	SWI5	1.24E-02	7.05E-03	Pic *et al.*, 2000
25	PHD1	SKN7	3.10E-05	1.42E-02	NA
26	SWI5	YAP5	1.84E-02	1.60E-02	NA
27	ACE2	NDD1	2.13E-02	6.29E-03	NA
28	SWI4	YAP1	1.89E-02	2.33E-02	NA
29	PDR1	RAP1	1.74E-07	2.39E-02	NA
30	MBP1	RME1	2.78E-02	5.06E-04	NA
31	ACE2	PHO4	1.81E-03	4.80E-02	NA
32	MBP1	PHO4	1.02E-02	4.80E-02	NA
33	PHO4	SKN7	4.80E-02	5.24E-04	NA
34	ACE2	SKN7	4.95E-02	1.30E-02	NA

Supplementary Table A27 $D(i,j,3)$, $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 3$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	HIR1	HIR2	1.88E-05	1.77E-06	Loy *et al.*, 1999
2	FHL1	RAP1	3.35E-05	9.62E-61	NA
3	NRG1	PHD1	5.44E-04	1.64E-05	NA
4	MCM1	SWI4	9.36E-04	4.27E-04	NA
5	FHL1	PDR1	1.39E-03	2.29E-08	NA
6	MBP1	MCM1	2.58E-03	6.11E-05	NA
7	FHL1	GAT3	3.61E-03	5.78E-09	NA
8	FKH2	MCM1	4.17E-03	4.92E-04	Spector *et al.*, 1997
9	MBP1	MSN4	4.82E-03	5.87E-04	NA
10	NDD1	SWI5	2.38E-03	7.80E-03	NA
11	RAP1	SFP1	9.65E-03	1.74E-03	NA
12	HAP2	HAP5	3.86E-04	1.01E-02	McNabb *et al.*, 1995
13	HAP3	HAP5	7.20E-04	1.01E-02	McNabb *et al.*, 1995
14	NDD1	SKN7	1.02E-02	3.24E-03	NA
15	MBP1	SKN7	1.10E-02	2.10E-07	Bouquin *et al.*, 1999
16	ARG80	ARG81	1.40E-02	4.37E-04	Mamnun *et al.*, 2002
17	ACE2	NDD1	1.62E-02	2.89E-05	NA
18	FKH1	NDD1	1.66E-02	1.63E-02	Kumar *et al.*, 2000
19	NRG1	SKN7	4.57E-04	1.75E-02	NA
20	SWI5	YAP1	1.93E-02	1.87E-02	NA
21	RAP1	YAP5	2.05E-02	1.48E-10	NA
22	SKN7	YAP1	3.94E-03	2.05E-02	Lee *et al.*, 1999
23	NDD1	RME1	1.63E-02	2.21E-02	NA
24	ACE2	MBP1	2.48E-02	2.53E-02	NA
25	ARG81	GCN4	4.58E-03	2.59E-02	NA
26	ACE2	FKH2	3.83E-02	1.10E-03	Pic *et al.*, 2000
27	MBP1	YAP1	4.73E-02	2.05E-02	NA

Supplementary Table A28 $D(i,j,3)$, $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 5$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	FHL1	RAP1	2.02E-20	4.40E-34	NA
2	PDR1	RAP1	2.00E-10	5.25E-10	NA
3	FHL1	PDR1	1.57E-07	7.06E-09	NA
4	ABF1	REB1	8.75E-05	3.12E-05	Schuller *et al.*, 1994
5	RAP1	YAP5	3.17E-04	1.04E-04	NA
6	HIR1	HIR2	4.64E-04	2.92E-04	Loy *et al.*, 1999
7	ACE2	NDD1	2.27E-03	7.95E-05	NA
8	FKH2	NDD1	3.49E-03	1.38E-03	Kumar *et al.*, 2000
9	MBP1	PHO4	2.51E-03	3.89E-03	NA
10	PHO4	SKN7	3.89E-03	2.62E-04	NA
11	ACE2	PHO4	3.93E-03	3.89E-03	NA
12	MET31	MET4	5.49E-03	4.25E-03	Blaiseau *et al.*, 1997
13	ACE2	YAP1	1.40E-03	5.99E-03	NA
14	SWI5	YAP1	1.50E-04	5.99E-03	NA
15	NDD1	SWI5	6.22E-03	1.60E-03	NA
16	MBP1	YAP1	7.00E-04	7.90E-03	NA
17	SKN7	YAP1	1.28E-04	7.90E-03	Lee *et al.*, 1999
18	NDD1	SWI4	4.01E-04	7.95E-03	NA
19	RME1	SKN7	5.01E-03	7.98E-03	NA
20	RME1	SWI5	5.01E-03	8.72E-03	NA
21	FKH2	YAP1	4.25E-04	8.72E-03	NA
22	ACE2	MBP1	4.78E-03	8.98E-03	NA
23	MBP1	SKN7	8.99E-03	5.94E-06	NA
24	NRG1	PHD1	9.35E-03	9.96E-03	NA
25	HAP2	HAP5	3.86E-04	1.01E-02	McNabb *et al.*, 1995
26	HAP3	HAP5	1.60E-03	1.01E-02	McNabb *et al.*, 1995
27	FHL1	GAT3	1.27E-02	1.83E-05	NA
28	RME1	SWI4	5.01E-03	1.38E-02	NA
29	FKH2	RME1	1.40E-02	5.01E-03	NA
30	ARG80	ARG81	1.40E-02	4.37E-04	Mamnun *et al.*, 2002
31	MBP1	MSN4	2.24E-02	2.82E-03	NA
32	ACE2	FKH2	2.30E-02	1.49E-02	Pic *et al.*, 2000
33	FKH2	PHO4	1.40E-02	2.74E-02	NA
34	NDD1	PHO4	7.04E-03	2.74E-02	NA
35	PHO4	RME1	2.74E-02	5.01E-03	NA
36	PHO4	SWI4	2.74E-02	1.38E-02	NA
37	PHO4	SWI5	2.74E-02	8.72E-03	NA
38	RAP1	SFP1	2.78E-02	3.08E-03	NA
39	ACE2	RME1	3.02E-02	5.01E-03	NA
40	MBP1	MCM1	3.86E-02	5.02E-03	NA
41	ABF1	MBP1	3.99E-02	7.63E-03	NA
42	ABF1	SWI6	3.99E-02	7.13E-03	NA
43	NDD1	YAP1	4.11E-04	4.56E-02	NA

Supplementary Table A29 $D(i,j,3)$, $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 10$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	PDR1	RAP1	1.71E-09	2.38E-07	NA
2	FKH2	NDD1	1.40E-06	2.12E-10	Kumar *et al.*, 2000
3	FHL1	PDR1	6.00E-06	3.91E-07	NA
4	ABF1	REB1	2.61E-04	4.60E-05	Schuller *et al.*, 1994
5	ACE2	FKH2	4.53E-04	1.05E-06	Pic *et al.*, 2000
6	FKH2	SWI5	9.55E-04	9.29E-04	Pic *et al.*, 2000
7	MBP1	SWI6	7.85E-09	1.36E-03	Koch *et al.*, 1993
8	HIR1	HIR2	5.95E-06	5.61E-03	Loy *et al.*, 1999
9	ACE2	SKN7	1.90E-02	7.58E-04	NA
10	ACE2	YAP1	1.39E-02	2.02E-02	NA
11	SWI5	YAP1	8.98E-04	2.02E-02	NA
12	MBP1	YAP1	2.40E-03	2.71E-02	NA
13	SKN7	YAP1	3.74E-03	2.71E-02	Lee *et al.*, 1999
14	RME1	SWI5	1.40E-02	2.82E-02	NA
15	FKH2	YAP1	1.74E-03	3.01E-02	NA
16	RME1	SWI4	1.40E-02	3.22E-02	NA
17	FKH2	RME1	3.28E-02	1.40E-02	NA

Supplementary Table A30 *in vivo* pull-down, $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 1$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	ABF1	REB1	1.14E-04	6.41E-05	Schuller *et al.*, 1994
2	FKH2	SWI5	2.14E-04	6.84E-04	Pic *et al.*, 2000
3	PHO4	SKN7	8.32E-04	2.23E-04	NA
4	MBP1	PHO4	1.54E-03	8.32E-04	NA
5	FKH2	SWI4	6.20E-08	2.52E-03	NA
6	HIR1	HIR2	6.40E-06	5.61E-03	Loy *et al.*, 1999
7	RAP1	YAP5	2.69E-03	6.21E-03	NA
8	FKH2	RME1	7.49E-03	1.01E-02	NA
9	NDD1	RME1	4.71E-03	1.01E-02	NA
10	RME1	SKN7	1.01E-02	1.10E-02	NA
11	ACE2	PHO4	1.18E-02	8.32E-04	NA
12	ACE2	SKN7	1.28E-02	2.20E-08	NA
13	RME1	SWI4	1.01E-02	1.59E-02	NA
14	RME1	SWI5	1.01E-02	1.63E-02	NA
15	FKH2	PHO4	7.49E-03	2.12E-02	NA
16	NDD1	PHO4	4.71E-03	2.12E-02	NA
17	PHO4	RME1	2.12E-02	1.01E-02	NA
18	PHO4	SWI4	2.12E-02	1.59E-02	NA
19	PHO4	SWI5	2.12E-02	1.63E-02	NA
20	MBP1	RME1	2.95E-02	1.01E-02	NA
21	HSF1	RAP1	3.13E-02	5.98E-03	NA
22	MBP1	SKN7	3.63E-02	3.61E-03	Bouquin *et al.*, 1999
23	DIG1	STE12	4.71E-02	2.06E-06	Bardwell *et al.*, 1998
24	FKH2	NDD1	4.81E-02	1.74E-06	Kumar *et al.*, 2000

Supplementary Table A31 *in vivo* pull-down, $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 2$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	ABF1	REB1	1.42E-04	8.12E-05	Schuller *et al.*, 1994
2	FKH2	SWI5	4.03E-04	9.29E-04	Pic *et al.*, 2000
3	PHO4	SKN7	1.18E-03	2.23E-04	NA
4	MBP1	PHO4	2.32E-03	1.18E-03	NA
5	FKH2	SWI4	1.64E-05	2.52E-03	NA
6	HIR1	HIR2	6.40E-06	5.61E-03	Loy *et al.*, 1999
7	RAP1	YAP5	4.70E-03	6.21E-03	NA
8	NDD1	RME1	5.22E-03	1.01E-02	NA
9	FKH2	RME1	1.07E-02	1.01E-02	NA
10	RME1	SKN7	1.01E-02	1.10E-02	NA
11	ACE2	PHO4	1.18E-02	1.18E-03	NA
12	ACE2	SKN7	1.28E-02	2.20E-08	NA
13	RME1	SWI4	1.01E-02	1.59E-02	NA
14	RME1	SWI5	1.01E-02	1.94E-02	NA
15	FKH2	PHO4	1.07E-02	2.61E-02	NA
16	NDD1	PHO4	5.22E-03	2.61E-02	NA
17	PHO4	RME1	2.61E-02	1.01E-02	NA
18	PHO4	SWI4	2.61E-02	1.59E-02	NA
19	PHO4	SWI5	2.61E-02	1.94E-02	NA
20	HSF1	RAP1	3.13E-02	7.09E-03	NA
21	FKH2	NDD1	3.52E-02	1.63E-03	Kumar *et al.*, 2000
22	MBP1	RME1	3.77E-02	1.01E-02	NA
23	DIG1	STE12	4.71E-02	3.06E-06	Bardwell *et al.*, 1998

Supplementary Table A32 *in vivo* pull-down, $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 3$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	ABF1	REB1	1.64E-04	8.68E-05	Schuller *et al.*, 1994
2	FKH2	SWI5	4.03E-04	9.29E-04	Pic *et al.*, 2000
3	PHO4	SKN7	2.34E-03	3.74E-04	NA
4	MBP1	PHO4	2.96E-03	2.34E-03	NA
5	HIR1	HIR2	6.40E-06	5.61E-03	Loy *et al.*, 1999
6	MBP1	SWI6	4.42E-03	7.32E-03	Koch *et al.*, 1993
7	NDD1	RME1	5.22E-03	1.01E-02	NA
8	FKH2	RME1	1.07E-02	1.01E-02	NA
9	ACE2	PHO4	1.18E-02	2.34E-03	NA
10	ACE2	SKN7	1.28E-02	9.79E-07	NA
11	RME1	SKN7	1.01E-02	1.54E-02	NA
12	RAP1	YAP5	8.81E-03	1.64E-02	NA
13	RME1	SWI5	1.01E-02	1.94E-02	NA
14	RME1	SWI4	1.01E-02	2.04E-02	NA
15	FKH2	NDD1	3.52E-02	1.63E-03	Kumar *et al.*, 2000
16	FKH2	PHO4	1.07E-02	3.96E-02	NA
17	NDD1	PHO4	5.22E-03	3.96E-02	NA
18	PHO4	RME1	3.96E-02	1.01E-02	NA
19	PHO4	SWI4	3.96E-02	2.04E-02	NA
20	PHO4	SWI5	3.96E-02	1.94E-02	NA
21	FKH2	SWI4	1.64E-05	4.01E-02	NA
22	MBP1	RME1	4.34E-02	1.01E-02	NA
23	DIG1	STE12	4.71E-02	3.06E-06	Bardwell *et al.*, 1998
24	HSF1	RAP1	4.83E-02	8.67E-03	NA

Supplementary Table A33 *in vivo* pull-down, $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 10$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	ABF1	REB1	5.13E-04	1.17E-04	Schuller *et al.*, 1994
2	DIG1	STE12	1.50E-03	2.33E-03	Bardwell *et al.*, 1998
3	FKH2	SWI5	4.04E-03	4.39E-02	Pic *et al.*, 2000

Supplementary Table A34 Integration of localization data, $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 2$, $I_{\text{loc}} = 0.01$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	HIR1	HIR2	1.60E-07	6.97E-08	Loy *et al.*, 1999
2	NRG1	PHD1	5.27E-07	2.66E-07	NA
3	NRG1	YAP6	5.48E-15	1.09E-06	NA
4	MCM1	SWI4	2.33E-06	7.07E-06	NA
5	MBP1	SKN7	4.17E-05	1.52E-09	Bouquin *et al.*, 1999
6	FKH2	MCM1	5.46E-05	5.01E-08	Spector *et al.*, 1997
7	FHL1	PDR1	4.75E-04	1.77E-12	NA
8	MBP1	MSN4	1.16E-03	1.05E-03	NA
9	MET31	MET4	1.36E-03	1.29E-03	Blaiseau *et al.*, 1997
10	ACE2	SKN7	1.89E-03	2.18E-04	NA
11	MBP1	MCM1	2.08E-03	6.18E-06	NA
12	FKH2	SWI5	2.92E-03	2.16E-03	Pic *et al.*, 2000
13	FKH2	SWI6	2.72E-03	3.54E-03	NA
14	ACE2	YAP1	2.37E-03	3.89E-03	NA
15	SWI5	YAP1	1.46E-03	3.89E-03	NA
16	SKN7	YAP1	1.57E-03	3.92E-03	Lee *et al.*, 1999
17	FHL1	GAT3	4.83E-03	1.38E-08	NA
18	NDD1	YAP1	2.88E-04	5.44E-03	NA
19	MBP1	YAP1	5.53E-03	3.92E-03	NA
20	HAP2	HAP5	4.40E-03	5.56E-03	McNabb *et al.*, 1995
21	HAP3	HAP5	5.66E-03	5.56E-03	McNabb *et al.*, 1995
22	ACE2	FKH2	5.80E-03	4.73E-03	Pic *et al.*, 2000
23	MAC1	ROX1	5.93E-03	2.87E-06	NA
24	FHL1	RAP1	6.83E-03	2.46E-43	NA
25	NDD1	RME1	3.16E-03	7.05E-03	NA
26	CIN5	RAP1	3.93E-03	8.04E-03	NA
27	SWI5	YAP5	8.30E-03	7.48E-03	NA
28	DIG1	STE12	9.38E-03	4.10E-09	Bardwell *et al.*, 1998
29	FKH2	MBP1	5.92E-06	9.95E-03	NA
30	NDD1	SKN7	1.13E-03	1.01E-02	NA
31	ACE2	NDD1	2.04E-02	1.14E-03	NA
32	SKN7	YAP6	2.25E-02	1.10E-02	NA
33	ABF1	REB1	2.26E-02	1.77E-02	Schuller *et al.*, 1994
34	IME4	YAP5	2.42E-02	2.15E-02	NA
35	NRG1	RLM1	3.69E-02	2.30E-02	NA
36	PDR1	RAP1	4.29E-06	4.24E-02	NA
37	CIN5	NRG1	4.59E-02	4.11E-05	NA

Supplementary Table A35 Integration of function data, $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 2$, $I_{\text{fnc}} = 0.01$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	FHL1	RAP1	1.36E-10	3.38E-67	NA
2	HIR1	HIR2	6.40E-08	5.99E-09	Loy *et al.*, 1999
3	MCM1	SWI4	2.00E-06	1.60E-05	NA
4	ACE2	FKH2	6.27E-07	2.63E-05	Pic *et al.*, 2000
5	MBP1	SKN7	1.59E-04	5.08E-10	Bouquin *et al.*, 1999
6	FKH2	MCM1	2.50E-04	5.03E-12	Spector *et al.*, 1997
7	FHL1	PDR1	5.06E-04	2.74E-11	NA
8	HAP2	HAP5	1.54E-04	5.69E-04	McNabb *et al.*, 1995
9	HAP3	HAP5	2.27E-04	5.69E-04	McNabb *et al.*, 1995
10	MBP1	MSN4	9.62E-04	1.20E-03	NA
11	NRG1	PHD1	1.31E-03	1.44E-04	NA
12	FKH2	MBP1	2.38E-04	1.80E-03	NA
13	NDD1	YAP1	3.62E-04	2.06E-03	NA
14	NDD1	RME1	2.42E-03	1.99E-03	NA
15	FHL1	GAT3	2.43E-03	8.53E-09	NA
16	SWI5	YAP1	2.46E-03	2.68E-03	NA
17	ACE2	YAP1	3.31E-03	2.39E-03	NA
18	MBP1	MCM1	3.61E-03	5.43E-06	NA
19	MCM1	SWI6	1.31E-04	4.17E-03	NA
20	MBP1	YAP1	5.89E-03	2.29E-03	NA
21	NRG1	YAP6	1.49E-06	6.02E-03	NA
22	MAC1	ROX1	6.71E-03	1.40E-06	NA
23	ACE2	SKN7	6.73E-03	4.80E-06	NA
24	SWI5	YAP5	9.69E-03	6.82E-03	NA
25	RAP1	YAP5	1.15E-02	2.67E-14	NA
26	NDD1	SKN7	1.21E-02	6.88E-03	NA
27	HAP2	HAP3	1.42E-02	1.29E-02	McNabb *et al.*, 1995
28	FKH2	SWI5	2.21E-02	7.13E-03	Pic *et al.*, 2000
29	ABF1	REB1	2.28E-02	1.96E-02	Schuller *et al.*, 1994
30	ACE2	NDD1	2.57E-02	4.39E-03	NA
31	NRG1	RLM1	2.60E-02	1.95E-02	NA
32	ACE2	SWI5	2.86E-02	9.03E-03	Zhu *et al.*, 2000
33	ACE2	SWI4	3.78E-02	7.83E-03	NA
34	GCR1	GCR2	4.21E-02	8.72E-03	Turkel and Bisson, 1999; Uemura *et al.*, 1997
35	CIN5	RAP1	2.92E-03	4.50E-02	NA
36	ACE2	PHO4	6.17E-03	4.73E-02	NA
37	MBP1	PHO4	1.23E-02	4.73E-02	NA

Supplementary Table A36 Integration of localization and function data, $P_B < 0.001$, $O_{\min} = 3$, $I_{\min} = 2$, $I_{\text{loc}} = 0.01$, $I_{\text{fnc}} = 0.01$, $P_{mw} \leq 0.05$

#	TF1	TF2	P_{mw}(vs. TF1)	P_{mw}(vs. TF2)	L.E.
1	HIR1	HIR2	1.43E-07	6.97E-08	Loy *et al.*, 1999
2	NRG1	PHD1	4.99E-07	1.62E-07	NA
3	NRG1	YAP6	8.08E-14	1.17E-06	NA
4	MCM1	SWI4	1.99E-06	7.27E-06	NA
5	FKH2	MCM1	3.20E-05	4.55E-08	Spector *et al.*, 1997
6	MBP1	SKN7	3.54E-05	1.46E-09	Bouquin *et al.*, 1999
7	FHL1	PDR1	6.15E-04	3.63E-11	NA
8	ACE2	SKN7	6.55E-04	1.01E-04	NA
9	MBP1	MSN4	9.59E-04	8.21E-04	NA
10	MET31	MET4	1.17E-03	1.20E-03	Blaiseau *et al.*, 1997
11	MBP1	MCM1	1.57E-03	5.03E-06	NA
12	ACE2	YAP1	1.71E-03	2.65E-03	NA
13	SWI5	YAP1	9.24E-04	2.76E-03	NA
14	SKN7	YAP1	1.13E-03	2.84E-03	Lee *et al.*, 1999
15	NDD1	YAP1	1.56E-04	2.84E-03	NA
16	FKH2	SWI6	1.97E-03	3.41E-03	NA
17	FKH2	SWI5	3.71E-03	2.44E-03	Pic *et al.*, 2000
18	FHL1	GAT3	3.99E-03	8.94E-09	NA
19	MBP1	YAP1	4.40E-03	2.97E-03	NA
20	FHL1	RAP1	4.97E-03	4.31E-50	NA
21	ACE2	FKH2	5.00E-03	3.89E-03	Pic *et al.*, 2000
22	MAC1	ROX1	5.10E-03	3.91E-06	NA
23	NDD1	RME1	2.78E-03	5.37E-03	NA
24	HAP2	HAP5	4.40E-03	5.56E-03	McNabb *et al.*, 1995
25	HAP3	HAP5	5.66E-03	5.56E-03	McNabb *et al.*, 1995
26	CIN5	RAP1	3.68E-03	7.73E-03	NA
27	SWI5	YAP5	8.08E-03	6.69E-03	NA
28	FKH2	MBP1	4.90E-06	8.74E-03	NA
29	ACE2	NDD1	1.22E-02	9.29E-04	NA
30	NDD1	SKN7	1.75E-03	1.55E-02	NA
31	DIG1	STE12	1.70E-02	1.04E-08	Bardwell *et al.*, 1998
32	ABF1	REB1	1.74E-02	1.52E-02	Schuller *et al.*, 1994
33	SKN7	YAP6	2.41E-02	1.28E-02	NA
34	NRG1	RLM1	3.63E-02	2.30E-02	NA
35	PDR1	RAP1	4.05E-06	3.99E-02	NA
36	CBF1	MET4	1.28E-02	4.75E-02	Wheeler *et al.*, 2003
37	CIN5	NRG1	4.84E-02	8.08E-05	NA

Supplementary Table A37 Positive pairs in the ADR dataset

	chemical compound	NIST ID of compound	adrenergic receptor	UniProt ID of receptor	antagonist/agonist
1	ergotamine	248068	ADRA1A	P35348	antatonist
2	ergotamine	248068	ADRA1B	P35368	antatonist
3	ergotamine	248068	ADRA1D	P25100	antatonist
4	ergotamine	248068	ADRA2A	P08913	antatonist
5	ergotamine	248068	ADRA2B	P18089	antatonist
6	ergotamine	248068	ADRA2C	P18825	antatonist
7	metazocine	298730	ADRA1A	P35348	antatonist
8	metazocine	298730	ADRA1B	P35368	antatonist
9	metazocine	298730	ADRA1D	P25100	antatonist
10	dibenzyline	290787	ADRA1A	P35348	antatonist
11	dibenzyline	290787	ADRA1B	P35368	antatonist
12	dibenzyline	290787	ADRA1D	P25100	antatonist
13	dibenzyline	290787	ADRA2A	P08913	antatonist
14	dibenzyline	290787	ADRA2B	P18089	antatonist
15	dibenzyline	290787	ADRA2C	P18825	antatonist
16	regitine	248099	ADRA1A	P35348	antatonist
17	regitine	248099	ADRA1B	P35368	antatonist
18	regitine	248099	ADRA1D	P25100	antatonist
19	regitine	248099	ADRA2A	P08913	antatonist
20	regitine	248099	ADRA2B	P18089	antatonist
21	regitine	248099	ADRA2C	P18825	antatonist
22	prazosin	107105	ADRA1A	P35348	antatonist
23	prazosin	107105	ADRA1B	P35368	antatonist
24	prazosin	107105	ADRA1D	P25100	antatonist
25	terazosin	121255	ADRA1A	P35348	antatonist
26	terazosin	121255	ADRA1B	P35368	antatonist
27	terazosin	121255	ADRA1D	P25100	antatonist
28	tolazoline	245997	ADRA1A	P35348	antatonist
29	tolazoline	245997	ADRA1B	P35368	antatonist
30	tolazoline	245997	ADRA1D	P25100	antatonist
31	tolazoline	245997	ADRA2A	P08913	antatonist
32	tolazoline	245997	ADRA2B	P18089	antatonist
33	tolazoline	245997	ADRA2C	P18825	antatonist
34	yohimbine	247969	ADRA2A	P08913	antatonist
35	yohimbine	247969	ADRA2B	P18089	antatonist
36	yohimbine	247969	ADRA2C	P18825	antatonist
37	acebutolol	120460	ADRB1	P08588	antatonist
38	alprenolol	42318	ADRB1	P08588	antatonist
39	alprenolol	42318	ADRB2	P07550	antatonist
40	atenolol	107108	ADRB1	P08588	antatonist
41	betaxolol	121082	ADRB1	P08588	antatonist
42	bisoprolol	159236	ADRB1	P08588	antatonist
43	bupranolol	190477	ADRB1	P08588	antatonist
44	carteolol	159278	ADRB1	P08588	antatonist
45	carteolol	159278	ADRB2	P07550	antatonist
46	levobunolol	121213	ADRB1	P08588	antatonist
47	levobunolol	121213	ADRB2	P07550	antatonist
48	metoprolol	246551	ADRB1	P08588	antatonist
49	metoprolol	246551	ADRB2	P07550	antatonist
50	nadolol	75219	ADRB1	P08588	antatonist
51	nadolol	75219	ADRB2	P07550	antatonist
52	oxprenolol	107107	ADRB1	P08588	antatonist
53	oxprenolol	107107	ADRB2	P07550	antatonist
54	pindolol	235693	ADRB1	P08588	antatonist
55	pindolol	235693	ADRB2	P07550	antatonist
56	practolol	25049	ADRB1	P08588	antatonist
57	propranolol	158021	ADRB1	P08588	antatonist
58	propranolol	158021	ADRB2	P07550	antatonist
59	sotalol	246587	ADRB1	P08588	antatonist
60	sotalol	246587	ADRB2	P07550	antatonist
61	timolol	246889	ADRB1	P08588	antatonist
62	timolol	246889	ADRB2	P07550	antatonist
63	toliprolol	188192	ADRB1	P08588	antatonist
64	toliprolol	188192	ADRB2	P07550	antatonist
65	labetalol	248725	ADRB1	P08588	antatonist
66	labetalol	248725	ADRB2	P07550	antatonist
67	labetalol	248725	ADRA2A	P08913	antatonist
68	labetalol	248725	ADRA2B	P18089	antatonist
69	labetalol	248725	ADRA2C	P18825	antatonist

continued on next page.

continued from previous page.

	chemical compound	NIST ID of compound	adrenergic receptor	UniProt ID of receptor	antagonist/agonist
70	clonidine	233859	ADRA2A	P08913	agonist
71	clonidine	233859	ADRA2B	P18089	agonist
72	clonidine	233859	ADRA2C	P18825	agonist
73	guanabenz	248361	ADRA2A	P08913	agonist
74	guanabenz	248361	ADRA2B	P18089	agonist
75	guanabenz	248361	ADRA2C	P18825	agonist
76	levarterenol	246031	ADRA1A	P35348	agonist
77	levarterenol	246031	ADRA1B	P35368	agonist
78	levarterenol	246031	ADRA1D	P25100	agonist
79	levarterenol	246031	ADRA2A	P08913	agonist
80	levarterenol	246031	ADRA2B	P18089	agonist
81	levarterenol	246031	ADRA2C	P18825	agonist
82	levarterenol	246031	ADRB1	P08588	agonist
83	metaraminol	248431	ADRA1A	P35348	agonist
84	metaraminol	248431	ADRA1B	P35368	agonist
85	metaraminol	248431	ADRA1D	P25100	agonist
86	metaraminol	248431	ADRA2A	P08913	agonist
87	metaraminol	248431	ADRA2B	P18089	agonist
88	metaraminol	248431	ADRA2C	P18825	agonist
89	methoxamine	246208	ADRA1A	P35348	agonist
90	methoxamine	246208	ADRA1B	P35368	agonist
91	methoxamine	246208	ADRA1D	P25100	agonist
92	oxymetazoline	113855	ADRA2A	P08913	agonist
93	oxymetazoline	113855	ADRA2B	P18089	agonist
94	oxymetazoline	113855	ADRA2C	P18825	agonist
95	phenylephrine	232105	ADRA1A	P35348	agonist
96	phenylephrine	232105	ADRA1B	P35368	agonist
97	phenylephrine	232105	ADRA1D	P25100	agonist
98	tizanidine	162088	ADRA2A	P08913	agonist
99	tizanidine	162088	ADRA2B	P18089	agonist
100	tizanidine	162088	ADRA2C	P18825	agonist
101	albuterol	298757	ADRB2	P07550	agonist
102	bitolterol	248773	ADRB2	P07550	agonist
103	clenbuterol	74553	ADRB2	P07550	agonist
104	clorprenaline	298621	ADRB2	P07550	agonist
105	dobutamine	235669	ADRB1	P08588	agonist
106	isoetharine	247952	ADRB2	P07550	agonist
107	isoprenaline	231638	ADRB1	P08588	agonist
108	isoprenaline	231638	ADRB2	P07550	agonist
109	metaproterenol	248040	ADRB2	P07550	agonist
110	ritodrine	248145	ADRB2	P07550	agonist
111	terbutaline	298780	ADRB2	P07550	agonist
112	dopamine	228609	ADRA1A	P35348	agonist
113	dopamine	228609	ADRA1B	P35368	agonist
114	dopamine	228609	ADRA1D	P25100	agonist
115	dopamine	228609	ADRA2A	P08913	agonist
116	dopamine	228609	ADRA2B	P18089	agonist
117	dopamine	228609	ADRA2C	P18825	agonist
118	dopamine	228609	ADRB1	P08588	agonist
119	dopamine	228609	ADRB2	P07550	agonist
120	ephedrine	113823	ADRA1A	P35348	agonist
121	ephedrine	113823	ADRA1B	P35368	agonist
122	ephedrine	113823	ADRA1D	P25100	agonist
123	ephedrine	113823	ADRA2A	P08913	agonist
124	ephedrine	113823	ADRA2B	P18089	agonist
125	ephedrine	113823	ADRA2C	P18825	agonist
126	ephedrine	113823	ADRB1	P08588	agonist
127	ephedrine	113823	ADRB2	P07550	agonist
128	epinephrine	24257	ADRA1A	P35348	agonist
129	epinephrine	24257	ADRA1B	P35368	agonist
130	epinephrine	24257	ADRA1D	P25100	agonist
131	epinephrine	24257	ADRA2A	P08913	agonist
132	epinephrine	24257	ADRA2B	P18089	agonist
133	epinephrine	24257	ADRA2C	P18825	agonist
134	epinephrine	24257	ADRB1	P08588	agonist
135	epinephrine	24257	ADRB2	P07550	agonist
136	mephentermine	113865	ADRA1A	P35348	agonist
137	mephentermine	113865	ADRA1B	P35368	agonist
138	mephentermine	113865	ADRA1D	P25100	agonist

continued on next page.

continued from previous page.

	chemical compound	NIST ID of compound	adrenergic receptor	UniProt ID of receptor	antagonist/agonist
139	mephentermine	113865	ADRA2A	P08913	agonist
140	mephentermine	113865	ADRA2B	P18089	agonist
141	mephentermine	113865	ADRA2C	P18825	agonist
142	mephentermine	113865	ADRB1	P08588	agonist

continued on next page.

Supplementary Table A38 Positive pairs in the DrugBank dataset

	chemical compound	NIST ID of compound	UniProt ID of protein
1	Acebutolol	120460	P08588
2	Acetaminophen	229798	P23319
3	Acetaminophen	229798	P35354
4	Acetazolamide	239163	P00915
5	Acetohexamide	247668	P48048
6	Acetohydroxamic Acid	231100	P18314
7	Acetophenazine	247953	P14416
8	Acetophenazine	247953	P21728
9	Acitretin	141590	P10276
10	Acitretin	141590	P10826
11	Acitretin	141590	P13631
12	Acitretin	141590	P19793
13	Acitretin	141590	P22932
14	Acitretin	141590	P28702
15	Acitretin	141590	P48443
16	Acyclovir	248607	P04293
17	Acyclovir	248607	P09252
18	Adenosine	227956	Q08828
19	Albendazole	256773	P50719
20	Albuterol	298757	P07550
21	Alclometasone	235695	P04083
22	Alclometasone	235695	P08185
23	Alfentanil	248171	P35372
24	Allopurinol	230398	P47989
25	Alprazolam	250574	P14867
26	Alprazolam	250574	P30536
27	Alprazolam	250574	P31644
28	Alprazolam	250574	P34903
29	Alprazolam	250574	P47869
30	Alprazolam	250574	P48169
31	Alprazolam	250574	Q16445
32	Alprenolol	42318	P07550
33	Alprenolol	42318	P08588
34	Amantadine	232354	P10920
35	Amantadine	232354	P14416
36	Amantadine	232354	P21728
37	Amcinonide	248609	P04083
38	Amcinonide	248609	P04150
39	Amiloride	254340	P19634
40	Amiloride	254340	P19801
41	Amiloride	254340	P37088
42	Amiloride	254340	P51168
43	Amiloride	254340	P51170
44	Amiloride	254340	P51172
45	Amiloride	254340	P78348
46	Amiloride	254340	Q16515
47	Aminocaproic Acid	298481	P00747
48	Aminoglutethimide	247719	P11511
49	Aminophylline	257605	P33765
50	Aminophylline	257605	Q14432
51	Aminosalicylic Acid	228370	P64143
52	Amiodarone	120482	P08588
53	Amiodarone	120482	P35348
54	Amitriptyline	42327	P23975

continued on next page.

continued from previous page.

	chemical compound	NIST ID of compound	UniProt ID of protein
55	Amitriptyline	42327	P31645
56	Amlodipine	247418	P00915
57	Amlodipine	247418	Q06432
58	Amodiaquine	298615	P69905
59	Amoxapine	247984	P23975
60	Amoxapine	247984	P31645
61	Amphetamine	235438	P08913
62	Amphetamine	235438	P35348
63	Amsacrine	131642	P11388
64	Anisindione	120491	P38435
65	Apomorphine	248048	P14416
66	Apomorphine	248048	P21728
67	Apomorphine	248048	P21917
68	Aspirin	221215	P23319
69	Aspirin	221215	P35354
70	Astemizole	292149	P35367
71	Atenolol	107108	P08588
72	Atovaquone	247458	P43264
73	Atropine	221030	P08172
74	Atropine	221030	P08173
75	Atropine	221030	P08912
76	Atropine	221030	P11229
77	Atropine	221030	P20309
78	Azatadine	248317	P35367
79	Azelaic Acid	113078	P00440
80	Azelaic Acid	113078	P00582
81	Azelaic Acid	113078	P31213
82	Azelaic Acid	113078	P66011
83	Baclofen	290937	Q9UBS5
84	Beclomethasone	247723	P04083
85	Beclomethasone	247723	P08185
86	Bendroflumethiazide	247724	P00915
87	Bendroflumethiazide	247724	P00918
88	Bendroflumethiazide	247724	P22748
89	Bendroflumethiazide	247724	P55017
90	Bendroflumethiazide	247724	Q12791
91	Bentiromide	248616	P04746
92	Bentiromide	248616	P16233
93	Benzocaine	250591	Q9Y5Y9
94	Benzonatate	247726	Q14524
95	Benzphetamine	250592	P08913
96	Benzphetamine	250592	P35348
97	Benzquinamide	247108	P08172
98	Benzquinamide	247108	P08173
99	Benzquinamide	247108	P08912
100	Benzquinamide	247108	P11229
101	Benzquinamide	247108	P20309
102	Benzquinamide	247108	P35367
103	Benzthiazide	247728	P00915
104	Benzthiazide	247728	P00918
105	Benzthiazide	247728	P22748
106	Benzthiazide	247728	P55017
107	Benzthiazide	247728	Q12791
108	Benztropine	246818	P11229
109	Betamethasone	74535	P04083
110	Betamethasone	74535	P04150
111	Betaxolol	121082	P08588
112	Betazole	298618	P25021
113	Biperiden	247729	P11229
114	Biperiden	247729	Q15822
115	Bisoprolol	159236	P07550
116	Bisoprolol	159236	P08588
117	Bitolterol Mesylate	248773	P07550
118	Bromocriptine	248047	P14416
119	Bromodiphenhydramine	247732	P35367
120	Brompheniramine	250546	P35367
121	Buclizine	248626	P11229
122	Buclizine	248626	P35367
123	Budesonide	248257	P04150
124	Bumetanide	248772	P05023
125	Bumetanide	248772	Q13621
126	Bupivacaine	159229	P34995
127	Bupivacaine	159229	Q9Y5Y9

continued on next page.

continued from previous page.

	chemical compound	NIST ID of compound	UniProt ID of protein
128	Buprenorphine	248591	P35372
129	Buprenorphine	248591	P41145
130	Bupropion	248891	P23975
131	Bupropion	248891	Q01959
132	Buspirone	248867	P08908
133	Buspirone	248867	P14416
134	Butabarbital	250699	P14867
135	Butalbital	10652	P14867
136	Butoconazole	248868	P50859
137	Butorphanol Tartrate	247735	P35372
138	Butorphanol Tartrate	247735	P41145
139	Caffeine	290714	P30542
140	Caffeine	290714	Q07343
141	Calcitriol	248752	O15528
142	Calcitriol	248752	P11473
143	Captopril	250717	P22966
144	Carbachol	230651	P08172
145	Carbachol	230651	P11229
146	Carbachol	230651	Q15822
147	Carbamazepine	236284	Q14524
148	Carbidopa	298626	P20711
149	Carbimazole	107111	P07202
150	Carbinoxamine	58729	P11229
151	Carbinoxamine	58729	P35367
152	Carmustine	248168	P00390
153	Carteolol	159278	P07550
154	Carteolol	159278	P08588
155	Cefaclor	235663	Q8XJ01
156	Cephalexin	235673	Q8XJ01
157	Chloramphenicol	52747	P0A7J3
158	Chlordiazepoxide	23260	P14867
159	Chlormezanone	233880	P30536
160	Chloroprocaine	137198	P46098
161	Chloroprocaine	137198	Q01959
162	Chloroprocaine	137198	Q8TCU5
163	Chloroprocaine	137198	Q9GZZ6
164	Chloroprocaine	137198	Q9Y5Y9
165	Chloroquine	42361	P69905
166	Chlorothiazide	244641	P00915
167	Chlorothiazide	244641	P00918
168	Chlorothiazide	244641	P22748
169	Chlorothiazide	244641	P55017
170	Chlorothiazide	244641	Q12791
171	Chlorotrianisene	248055	P03372
172	Chlorpheniramine	250548	P35367
173	Chlorpromazine	250594	P14416
174	Chlorpromazine	250594	P28223
175	Chlorpropamide	233879	P48048
176	Chlorprothixene	247780	P14416
177	Chlorprothixene	247780	P21728
178	Chlorthalidone	246980	Q13621
179	Chlorzoxazone	239377	Q12791
180	Cholecalciferol	248815	P11473
181	Cholecalciferol	248815	Q02318
182	Cholecalciferol	248815	Q6VVX0
183	Ciclopirox	248393	P16050
184	Ciclopirox	248393	P23319
185	Ciclopirox	248393	P35354
186	Cimetidine	237089	P25021
187	Cinnarizine	158442	O00555
188	Cinnarizine	158442	P14416
189	Cinnarizine	158442	P35367
190	Ciprofloxacin	121210	P11388
191	Cisapride	121231	Q13639
192	Clemastine	247788	P35367
193	Clindamycin	247789	P0A7J3
194	Clobazam	107121	P14867
195	Clocortolone	248665	P04083
196	Clofazimine	298622	Q5L2G3
197	Clofibrate	245138	P06858
198	Clomifene	248400	P03372
199	Clomipramine	246870	P09211
200	Clomipramine	246870	P23975

continued on next page.

continued from previous page.

	chemical compound	NIST ID of compound	UniProt ID of protein
201	Clomipramine	246870	P31645
202	Clonazepam	250715	P14867
203	Clonidine	233859	P08913
204	Clotrimazole	191548	P10614
205	Clozapine	247212	P14416
206	Clozapine	247212	P21917
207	Clozapine	247212	P28223
208	Clozapine	247212	P35367
209	Clozapine	247212	Q9H3N8
210	Clozapine	247212	Q9NYX4
211	Cocaine	113834	P21728
212	Cocaine	113834	P23975
213	Cocaine	113834	P30531
214	Cocaine	113834	P31645
215	Cocaine	113834	P35462
216	Cocaine	113834	P41145
217	Cocaine	113834	Q01959
218	Cocaine	113834	Q14524
219	Cocaine	113834	Q9UI33
220	Cocaine	113834	Q9Y5Y9
221	Codeine	313075	P35372
222	Codeine	313075	P41143
223	Codeine	313075	P41145
224	Cyclizine	250597	P08172
225	Cyclizine	250597	P11229
226	Cyclizine	250597	P20309
227	Cyclizine	250597	P35367
228	Cyclobenzaprine	246605	P28223
229	Cyclopentolate	292189	P11229
230	Cycloserine	237050	P0A6BA
231	Cycloserine	237050	P0A6JB
232	Cyclothiazide	298640	P00915
233	Cyclothiazide	298640	P00918
234	Cyclothiazide	298640	P22748
235	Cyclothiazide	298640	P54710
236	Cyclothiazide	298640	Q12791
237	Cycrimine	248053	P11229
238	Cyproheptadine	250706	P28223
239	Cytarabine	254325	P06746
240	Dantrolene	247798	P21827
241	Dapsone	51808	P29251
242	Demeclocycline	17858	P0A7V8
243	Demeclocycline	17858	P0A7X3
244	Deserpidine	247162	P12821
245	Desflurane	308772	P03886
246	Desflurane	308772	P14867
247	Desflurane	308772	P23415
248	Desflurane	308772	P30049
249	Desflurane	308772	P42261
250	Desflurane	308772	P98194
251	Desflurane	308772	Q09470
252	Desipramine	250707	P07550
253	Desipramine	250707	P08172
254	Desipramine	250707	P08588
255	Desipramine	250707	P11229
256	Desipramine	250707	P23975
257	Desipramine	250707	P31645
258	Desipramine	250707	P35367
259	Desoximetasone	121216	P04083
260	Dexamethasone	235905	P04083
261	Dexamethasone	235905	P04150
262	Dexbrompheniramine	15054	P35367
263	Dexrazoxane	131624	P11388
264	Dextromethorphan	235679	P31645
265	Dextromethorphan	235679	Q15822
266	Dextromethorphan	235679	Q5T1J1
267	Dextromethorphan	235679	Q8TCU5
268	Dextromethorphan	235679	Q99720
269	Diazepam	113837	P14867
270	Diazepam	113837	P30536
271	Diazoxide	292038	P00915
272	Diazoxide	292038	P00918
273	Diazoxide	292038	P22748

continued on next page.

continued from previous page.

	chemical compound	NIST ID of compound	UniProt ID of protein
274	Diazoxide	292038	P55017
275	Diazoxide	292038	Q12791
276	Dichlorphenamide	247909	P00915
277	Diclofenac	158432	P23319
278	Diclofenac	158432	P35354
279	Dicloxacillin	247910	Q8XJ01
280	Dicumarol	246974	Q9BQB6
281	Dicyclomine	250600	P11229
282	Didanosine	122987	P03369
283	Dienestrol	248819	P03372
284	Diethylcarbamazine	135586	P00395
285	Diethylcarbamazine	135586	P16050
286	Diethylpropion	246177	P23975
287	Diethylpropion	246177	Q01959
288	Diethylstilbestrol	234131	P03372
289	Diflorasone	248673	P04083
290	Diflunisal	248674	P23319
291	Diflunisal	248674	P35354
292	Digoxin	120638	P05023
293	Dihydroergotamine	248068	P28221
294	Dihydroergotamine	248068	P28222
295	Dihydrotachysterol	33089	P11473
296	Diltiazem	247122	Q06432
297	Diphenhydramine	250549	P35367
298	Diphenidol	247917	P08172
299	Diphenidol	247917	P11229
300	Diphenidol	247917	P20309
301	Diphenoxylate	125715	P35372
302	Diphenylpyraline	250708	P35367
303	Dipyridamole	237441	Q9Y233
304	Disopyramide	159197	Q14524
305	Disulfiram	228650	P05091
306	Disulfiram	228650	P30536
307	Dobutamine	235669	P08588
308	Dopamine	228609	P08588
309	Dopamine	228609	P09172
310	Dopamine	228609	P21728
311	Doxepin	292030	P23975
312	Doxepin	292030	P31645
313	Doxepin	292030	P35367
314	Doxorubicin	131605	P11388
315	Doxylamine	250552	P11229
316	Doxylamine	250552	P35367
317	Dromostanolone	313018	P03372
318	Dromostanolone	313018	P04278
319	Dromostanolone	313018	P10275
320	Dromostanolone	313018	P16471
321	Droperidol	247071	P14416
322	Dyclonine	65070	Q9Y5Y9
323	Dyphylline	239172	Q07343
324	Econazole	193280	P10614
325	Encainide	121124	Q14524
326	Enflurane	187240	P03886
327	Enflurane	187240	P14867
328	Enflurane	187240	P23415
329	Enflurane	187240	P30049
330	Enflurane	187240	P42261
331	Enflurane	187240	P98194
332	Enflurane	187240	Q09470
333	Epinephrine	24257	P07550
334	Epinephrine	24257	P08588
335	Epinephrine	24257	P35348
336	Ergocalciferol	248164	O15528
337	Ergocalciferol	248164	P11473
338	Ergocalciferol	248164	Q6VVX0
339	Ergoloid Mesylate	226969	P08908
340	Ergotamine	248677	P28222
341	Ergotamine	248677	P35348
342	Estradiol	13192	P03372
343	Estramustine	248681	P03372
344	Estrone	42853	P03372
345	Ethacrynic acid	159259	Q13621
346	Ethanol	118507	P14867

continued on next page.

continued from previous page.

	chemical compound	NIST ID of compound	UniProt ID of protein
347	Ethanol	118507	P23415
348	Ethanol	118507	P23416
349	Ethanol	118507	Q8TCU5
350	Ethinyl Estradiol	234172	P03372
351	Ethopropazine	246857	P11229
352	Ethopropazine	246857	Q8TCU5
353	Ethosuximide	250694	O43497
354	Ethoxzolamide	292040	P00915
355	Ethynodiol Diacetate	292151	P03372
356	Ethynodiol Diacetate	292151	P06401
357	Etoposide	131616	P11388
358	Felbamate	247587	Q8TCU5
359	Felodipine	247476	P54289
360	Fenfluramine	250583	P28335
361	Fentanyl	250541	P35372
362	Flecainide	248228	Q14524
363	Floxuridine	237402	P04818
364	Flucytosine	238324	P04818
365	Fludarabine	131665	P09884
366	Fludarabine	131665	P23921
367	Fludrocortisone	247932	P08235
368	Flunisolide	120228	P08185
369	Flunisolide	120228	P47712
370	Fluocinolone Acetonide	247934	P08185
371	Fluocinolone Acetonide	247934	P47712
372	Fluocinonide	292183	P08185
373	Fluocinonide	292183	P47712
374	Fluorescein	231984	P02768
375	Fluorometholone	248151	P08185
376	Fluorometholone	248151	P47712
377	Fluorouracil	230123	P04818
378	Fluoxetine	250697	P31645
379	Fluoxymesterone	247938	P03372
380	Fluoxymesterone	247938	P04278
381	Fluoxymesterone	247938	P10275
382	Fluoxymesterone	247938	P16471
383	Flupenthixol	247134	P14416
384	Flupenthixol	247134	P21728
385	Flupenthixol	247134	P35348
386	Fluphenazine	120153	P14416
387	Flurandrenolide	248069	P08185
388	Flurandrenolide	248069	P47712
389	Flurazepam	159279	P14867
390	Flurbiprofen	248229	P23219
391	Flurbiprofen	248229	P35354
392	Flutamide	248839	P10275
393	Fluvoxamine	250692	P31645
394	Fomepizole	272994	P00325
395	Fomepizole	272994	P00326
396	Fomepizole	272994	P07327
397	Furosemide	232586	P00915
398	Furosemide	232586	P00918
399	Furosemide	232586	P05023
400	Furosemide	232586	P22748
401	Furosemide	232586	Q12791
402	Furosemide	232586	Q13621
403	Galantamine	111316	P22303
404	Gemfibrozil	248360	P06858
405	Glipizide	248683	P48048
406	Glyburide	247159	P48048
407	Glycopyrrolate	226952	P11229
408	Griseofulvin	238364	P10875
409	Griseofulvin	238364	P50719
410	Guanabenz	248361	P08913
411	Guanethidine	232402	P08913
412	Guanethidine	232402	P18089
413	Guanethidine	232402	P18825
414	Guanethidine	232402	P23975
415	Guanethidine	232402	P25100
416	Guanethidine	232402	P35348
417	Guanethidine	232402	P35368
418	Guanfacine	226992	P08913
419	Halazepam	247942	P14867

continued on next page.

continued from previous page.

	chemical compound	NIST ID of compound	UniProt ID of protein
420	Halazepam	247942	P30536
421	Halofantrine	247531	Q76NM6
422	Haloperidol	197517	P14416
423	Halothane	238098	O15554
424	Halothane	238098	P03886
425	Halothane	238098	P14867
426	Halothane	238098	P23415
427	Halothane	238098	P30049
428	Halothane	238098	P42261
429	Halothane	238098	P98194
430	Halothane	238098	Q5SQR9
431	Hesperetin	237072	O75908
432	Hesperetin	237072	P35610
433	Hesperetin	237072	P55157
434	Hexachlorophene	232074	P06149
435	Hexylcaine	122651	Q14524
436	Hexylcaine	122651	Q9Y5Y9
437	Hydrochlorothiazide	74591	P00915
438	Hydrochlorothiazide	74591	P00918
439	Hydrochlorothiazide	74591	P22748
440	Hydrochlorothiazide	74591	P55017
441	Hydrochlorothiazide	74591	Q12791
442	Hydrocodone	250556	P35372
443	Hydrocodone	250556	P41143
444	Hydrocodone	250556	P41145
445	Hydrocortisone	16228	P04083
446	Hydrocortisone	16228	P04150
447	Hydromorphone	250603	P35372
448	Hydromorphone	250603	P41143
449	Hydromorphone	250603	P41145
450	Hydroxyurea	247947	P23921
451	Hydroxyzine	250626	P35367
452	Hyoscyamine	233146	P08172
453	Hyoscyamine	233146	P11229
454	Ibuprofen	233882	P23319
455	Ibuprofen	233882	P35354
456	Idoxuridine	230829	P04293
457	Imipramine	250696	P23975
458	Imipramine	250696	P30542
459	Imipramine	250696	P31645
460	Indapamide	248685	P15382
461	Indapamide	248685	P51787
462	Indomethacin	107188	P23319
463	Indomethacin	107188	P35354
464	Ipratropium	121219	P08172
465	Ipratropium	121219	P11229
466	Isocarboxazid	250701	P21397
467	Isocarboxazid	250701	P27338
468	Isoetharine	247952	P08588
469	Isoflurane	163170	P03886
470	Isoflurane	163170	P14867
471	Isoflurane	163170	P23415
472	Isoflurane	163170	P30049
473	Isoflurane	163170	P42261
474	Isoflurane	163170	P98194
475	Isoflurane	163170	Q09470
476	Isoflurophate	226257	P06276
477	Isoniazid	228778	P0A5Y6
478	Isoniazid	228778	Q08129
479	Isoproterenol	246203	P07550
480	Isoproterenol	246203	P08588
481	Isosorbide Dinitrate	298706	P16066
482	Isradipine	121153	P54289
483	Ketamine	157938	Q8TCU5
484	Ketoprofen	248419	P23319
485	Ketoprofen	248419	P35354
486	Ketorolac	247570	P23319
487	Ketorolac	247570	P35354
488	Ketotifen Fumarate	247227	P35367
489	Ketotifen Fumarate	247227	Q13946
490	Ketotifen Fumarate	247227	Q9NP56
491	Labetalol	248725	P07550
492	Labetalol	248725	P08588

continued on next page.

continued from previous page.

	chemical compound	NIST ID of compound	UniProt ID of protein
493	Labetalol	248725	P35348
494	Labetalol	248725	P35368
495	Levallorphan	246652	P35372
496	Levallorphan	246652	Q15822
497	Levobunolol	121213	P07550
498	Levobunolol	121213	P08588
499	Levodopa	229075	P14416
500	Levodopa	229075	P21728
501	Levomethadyl Acetate	58788	P35372
502	Levorphanol	246478	P35372
503	Lidocaine	113841	Q14524
504	Lidocaine	113841	Q9Y5Y9
505	Lindane	122234	Q75NA5
506	Loperamide	247826	O00555
507	Loperamide	247826	O15399
508	Loperamide	247826	P35372
509	Loperamide	247826	P62158
510	Loperamide	247826	Q12879
511	Loperamide	247826	Q13224
512	Loperamide	247826	Q14957
513	Loratadine	247301	P35367
514	Lorazepam	250605	P30536
515	Lovastatin	121161	P04035
516	Loxapine	246942	P14416
517	Loxapine	246942	P28223
518	Malathion	118987	P07140
519	Maprotiline	246615	P23975
520	Maprotiline	246615	P35348
521	Maprotiline	246615	P35367
522	Marinol	313141	P21554
523	Mazindol	247692	P23975
524	Mazindol	247692	Q01959
525	Mebendazole	120797	P50719
526	Mecamylamine	248629	Q15822
527	Meclizine	247693	P35367
528	Medroxyprogesterone	248042	P03372
529	Medroxyprogesterone	248042	P06401
530	Mefenamic acid	298720	P23219
531	Mefenamic acid	298720	P35354
532	Mefloquine	247457	P69905
533	Megestrol	291988	P03372
534	Megestrol	291988	P06401
535	Melatonin	235344	P48039
536	Meperidine	250571	P41145
537	Meperidine	250571	Q8TCU5
538	Mephenytoin	113831	Q14524
539	Mepivacaine	113842	Q9Y5Y9
540	Mequitazine	120802	P35367
541	Mercaptopurine	230460	P00491
542	Mercaptopurine	230460	P00492
543	Mesalamine	229892	P23319
544	Mesalamine	229892	P35354
545	Mesoridazine	247076	P14416
546	Mesoridazine	247076	P28223
547	Metaproterenol	248040	P07550
548	Metaraminol	248431	P35348
549	Metformin	238414	Q9Y478
550	Methacycline	248038	P0A7J3
551	Methadone	195723	P35372
552	Methadone	195723	Q8TCU5
553	Methadone	195723	Q9GZZ6
554	Methantheline	193105	P11229
555	Metharbital	42422	P14867
556	Methazolamide	248035	P00915
557	Methdilazine	290506	P35367
558	Methimazole	236047	P07202
559	Methohexital	247803	P14867
560	Methohexital	247803	P17787
561	Methohexital	247803	P23415
562	Methohexital	247803	Q15822
563	Methotrexate	75950	P00374
564	Methoxamine	246208	P35348
565	Methoxyflurane	248779	P03886

continued on next page.

continued from previous page.

	chemical compound	NIST ID of compound	UniProt ID of protein
566	Methoxyflurane	248779	P14867
567	Methoxyflurane	248779	P23415
568	Methoxyflurane	248779	P30049
569	Methoxyflurane	248779	P42261
570	Methoxyflurane	248779	P98194
571	Methoxyflurane	248779	Q09470
572	Methyclothiazide	16156	P00915
573	Methyclothiazide	16156	P00918
574	Methyclothiazide	16156	P05023
575	Methyclothiazide	16156	P22748
576	Methyclothiazide	16156	Q12791
577	Methyclothiazide	16156	Q13621
578	Methyldopa	246206	P08913
579	Methylergonovine	248696	P21728
580	Methylphenidate	113829	P23975
581	Methylphenidate	113829	Q01959
582	Methylprednisolone	247062	P04083
583	Methylprednisolone	247062	P04150
584	Methyprylon	107182	P14867
585	Methysergide	248028	P28223
586	Metoclopramide	160689	P11229
587	Metoclopramide	160689	P14416
588	Metolazone	248026	Q13621
589	Metoprolol	246551	P08588
590	Metronidazole	236890	O30585
591	Metronidazole	236890	P29166
592	Metyrosine	248728	P07101
593	Mexiletine	248218	Q14524
594	Miconazole	107110	P50859
595	Midazolam	248238	P14867
596	Milrinone	120312	Q07343
597	Minaprine	248239	P14416
598	Minaprine	248239	P21728
599	Minaprine	248239	P28223
600	Minaprine	248239	P28335
601	Minaprine	248239	P31645
602	Minaprine	248239	P41595
603	Minocycline	248023	P0A7V8
604	Minocycline	248023	P0A7X3
605	Minoxidil	248022	P48048
606	Mirtazapine	247432	P08913
607	Mirtazapine	247432	P28223
608	Mirtazapine	247432	P28335
609	Mirtazapine	247432	P46098
610	Mitotane	119004	P10109
611	Mitotane	119004	P15538
612	Mitoxantrone	131658	P11388
613	Moclobemide	247280	P21397
614	Morphine	42472	P35372
615	Nabilone	248241	P21554
616	Nabilone	248241	P34972
617	Nabumetone	292131	P23319
618	Nabumetone	292131	P35354
619	Nadolol	75219	P07550
620	Nadolol	75219	P08588
621	Nafcillin	248018	Q8XJ01
622	Nalbuphine	248211	P35372
623	Nalbuphine	248211	P41143
624	Nalbuphine	248211	P41145
625	Naloxone	246940	P35372
626	Naltrexone	247882	P35372
627	Nandrolone Phenpropionate	248687	P10275
628	Naproxen	237147	P23319
629	Naproxen	237147	P35354
630	Nefazodone	247544	P23975
631	Nefazodone	247544	P28223
632	Nefazodone	247544	P31645
633	Nefazodone	247544	P35348
634	Niacin	233225	P40261
635	Niacin	233225	Q15274
636	Niacin	233225	Q8TDS4
637	Nicardipine	247266	Q13936
638	Nicotine	281629	Q15822

continued on next page.

continued from previous page.

	chemical compound	NIST ID of compound	UniProt ID of protein
639	Nifedipine	158530	P54289
640	Nimodipine	247285	Q06432
641	Nisoldipine	248260	O00555
642	Nitrendipine	120451	Q06432
643	Nitric Oxide	31	P16066
644	Nitrofurantoin	290698	P0A7R5
645	Nitrofurantoin	290698	P17117
646	Nitrofurazone	75963	P06715
647	Nitrofurazone	75963	P07003
648	Nitrofurazone	75963	P61889
649	Nitrofurazone	75963	P77390
650	Nitroglycerin	246295	P16066
651	Norethindrone	234126	P06401
652	Norfloxacin	148019	P0AES6
653	Norfloxacin	148019	P72525
654	Norgestrel	246859	P03372
655	Norgestrel	246859	P06401
656	Norgestrel	246859	P35348
657	Nortriptyline	247889	P23975
658	Nortriptyline	247889	P31645
659	Olanzapine	281161	P08172
660	Olanzapine	281161	P08173
661	Olanzapine	281161	P08912
662	Olanzapine	281161	P11229
663	Olanzapine	281161	P14416
664	Olanzapine	281161	P20309
665	Olanzapine	281161	P21728
666	Olanzapine	281161	P21917
667	Olanzapine	281161	P28223
668	Olanzapine	281161	P28335
669	Olanzapine	281161	P35367
670	Orphenadrine	290525	P35367
671	Orphenadrine	290525	Q8TCU5
672	Ouabain	128089	P05023
673	Oxandrolone	313196	P10275
674	Oxaprozin	72485	P35354
675	Oxazepam	250610	P14867
676	Oxybuprocaine	133995	Q9Y5Y9
677	Oxybutynin	248178	P11229
678	Oxycodone	250611	P35372
679	Oxycodone	250611	P41143
680	Oxycodone	250611	P41145
681	Oxymetazoline	113855	P08913
682	Oxymetazoline	113855	P35348
683	Oxymorphone	250612	P35372
684	Oxyphencyclimine	248139	P08172
685	Oxyphencyclimine	248139	P11229
686	Oxyphencyclimine	248139	P20309
687	Oxytetracycline	248113	P0A7V8
688	Oxytetracycline	248113	P0A7X3
689	Paclitaxel	131606	P10415
690	Paclitaxel	131606	Q9H4B7
691	Papaverine	291096	Q07343
692	Paramethadione	247999	Q9P0X4
693	Paroxetine	247287	P31645
694	Penicillamine	231231	P04206
695	Penicillamine	231231	P29466
696	Penicillin V	256457	Q8XJ01
697	Pentazocine	125798	P35372
698	Pentobarbital	244543	P14867
699	Pentostatin	131639	P00813
700	Pentoxifylline	248838	Q07343
701	Pergolide	247504	P14416
702	Pergolide	247504	P21728
703	Perhexiline	120895	P23786
704	Perhexiline	120895	P50416
705	Perphenazine	241782	P14416
706	Perphenazine	241782	P21728
707	Perphenazine	241782	P35348
708	Phenacemide	52708	P35498
709	Phenelzine	121438	P21397
710	Phenformin	256456	Q9Y478
711	Phenindione	236131	P00734

continued on next page.

continued from previous page.

	chemical compound	NIST ID of compound	UniProt ID of protein
712	Phenmetrazine	113870	P21397
713	Phenmetrazine	113870	P23975
714	Phenmetrazine	113870	Q01959
715	Phenobarbitone	250543	P14867
716	Phenoxybenzamine	290787	P35348
717	Phenprocoumon	286539	Q9BQB6
718	Phentermine	250588	P23975
719	Phentermine	250588	Q01959
720	Phentolamine	248099	P08913
721	Phenylbutazone	14618	P35354
722	Phenylbutazone	14618	Q16647
723	Phenylephrine	232105	P35348
724	Phenylpropanolamine	250631	P08913
725	Phenylpropanolamine	250631	P35348
726	Phenytoin	232857	Q14524
727	Physostigmine	221137	P06276
728	Phytonadione	233846	P38435
729	Picrotoxin	298767	P14867
730	Picrotoxin	298767	P24046
731	Pilocarpine	248003	P11229
732	Pimozide	248701	O43497
733	Pimozide	248701	P14416
734	Pimozide	248701	P41143
735	Pindolol	235693	P07550
736	Pindolol	235693	P08588
737	Piperazine	229316	Q8ITG2
738	Pirenzepine	247006	P11229
739	Piroxicam	298765	P23219
740	Praziquantel	292191	P08515
741	Prazosin	107105	P35348
742	Prednisolone	234176	P08185
743	Prednisone	236144	P04083
744	Prednisone	236144	P04150
745	Prilocaine	113845	Q14524
746	Primaquine	92433	P69905
747	Primidone	244337	P14867
748	Probenecid	246675	O95820
749	Procainamide	125801	Q14524
750	Procaine	238047	P46098
751	Procaine	238047	Q01959
752	Procaine	238047	Q8TCU5
753	Procaine	238047	Q9GZZ6
754	Procaine	238047	Q9Y5Y9
755	Prochlorperazine	250629	P14416
756	Prochlorperazine	250629	P28223
757	Prochlorperazine	250629	P35367
758	Procyclidine	246695	P08172
759	Procyclidine	246695	P08173
760	Procyclidine	246695	P11229
761	Progesterone	62036	P03372
762	Progesterone	62036	P06401
763	Proguanil	298762	P13922
764	Promazine	292167	P08172
765	Promazine	292167	P08173
766	Promazine	292167	P08912
767	Promazine	292167	P11229
768	Promazine	292167	P14416
769	Promazine	292167	P20309
770	Promazine	292167	P21728
771	Promazine	292167	P21917
772	Promazine	292167	P25100
773	Promazine	292167	P28223
774	Promazine	292167	P28335
775	Promazine	292167	P35348
776	Promazine	292167	P35367
777	Promazine	292167	P35368
778	Promethazine	291997	P08172
779	Promethazine	291997	P08173
780	Promethazine	291997	P08912
781	Promethazine	291997	P11229
782	Promethazine	291997	P14416
783	Promethazine	291997	P20309
784	Promethazine	291997	P28223

continued on next page.

continued from previous page.

	chemical compound	NIST ID of compound	UniProt ID of protein
785	Promethazine	291997	P35348
786	Promethazine	291997	P35367
787	Propafenone	246996	Q12809
788	Propafenone	246996	Q14524
789	Proparacaine	248009	Q9Y5Y9
790	Propiomazine	246995	P14416
791	Propiomazine	246995	P28223
792	Propofol	227977	P14867
793	Propoxyphene	312352	P35372
794	Propoxyphene	312352	P41143
795	Propoxyphene	312352	P41145
796	Propranolol	158021	P07550
797	Propranolol	158021	P08588
798	Propylthiouracil	290664	P07202
799	Protriptyline	292033	P23975
800	Protriptyline	292033	P31645
801	Pseudoephedrine	246007	P07550
802	Pseudoephedrine	246007	P08913
803	Pseudoephedrine	246007	P35348
804	Pyrazinamide	118636	Q11195
805	Pyridostigmine	226928	P22303
806	Pyridoxine	228913	O00764
807	Pyridoxine	228913	P34896
808	Pyridoxine	228913	P35520
809	Pyridoxine	228913	Q96GD0
810	Pyridoxine	228913	Q9NVS9
811	Pyridoxine	228913	Q9Y617
812	Pyrimethamine	246424	P13922
813	Quinidine	125804	Q14524
814	Quinine	107180	P69905
815	Ranitidine	120443	P25021
816	Remoxipride	247514	P14416
817	Rescinnamine	247167	P12821
818	Reserpine	248144	Q05940
819	Ribavirin	213081	P16502
820	Ribavirin	213081	P20839
821	Ribavirin	213081	P22413
822	Ribavirin	213081	P26676
823	Ribavirin	213081	P49902
824	Ribavirin	213081	P55263
825	Riboflavin	120239	P30043
826	Riboflavin	120239	Q969G6
827	Rimantadine	273008	P10920
828	Ritodrine	248145	P07550
829	Scopolamine	234229	P11229
830	Secobarbital	232637	P14867
831	Selegiline	247283	P27338
832	Sertraline	247231	P31645
833	Sevoflurane	308798	P03886
834	Sevoflurane	308798	P14867
835	Sevoflurane	308798	P23415
836	Sevoflurane	308798	P30049
837	Sevoflurane	308798	P42261
838	Sevoflurane	308798	P98194
839	Sevoflurane	308798	Q09470
840	Sirolimus	131640	P41145
841	Sirolimus	131640	P62942
842	Spectinomycin	248735	P0A7S3
843	Spironolactone	153009	P30556
844	Sufentanil	248535	P35372
845	Sulfacetamide	242149	P05041
846	Sulfadiazine	248149	Q08210
847	Sulfamethizole	236955	P0AC13
848	Sulfamethoxazole	236956	P08192
849	Sulfametopyrazine	120977	Q08210
850	Sulfanilamide	228509	P37254
851	Sulfapyridine	232444	P0C002
852	Sulfasalazine	248712	P24752
853	Sulfasalazine	248712	Q9UPY5
854	Sulfinpyrazone	107193	P33527
855	Sulfisoxazole	242145	Q08210
856	Sulindac	215720	P35354
857	Sulpiride	248255	P14416

continued on next page.

continued from previous page.

	chemical compound	NIST ID of compound	UniProt ID of protein
858	Sumatriptan	247581	P28221
859	Sumatriptan	247581	P28222
860	Suprofen	159496	P23319
861	Suprofen	159496	P35354
862	Tacrine	291000	P22303
863	Tamoxifen	248059	P03372
864	Temazepam	248966	P30536
865	Teniposide	131633	P11388
866	Terazosin	121255	P35348
867	Terazosin	121255	P35368
868	Terbutaline	298780	P07550
869	Terfenadine	248740	P35367
870	Testolactone	247866	P11511
871	Testosterone	62026	P10275
872	Tetracycline	153011	P0A7V8
873	Tetracycline	153011	P0A7X3
874	Thalidomide	291688	P01375
875	Theophylline	118946	P29275
876	Theophylline	118946	Q07343
877	Theophylline	118946	Q14432
878	Thiabendazole	118975	P00363
879	Thiamine	228577	O60779
880	Thiamine	228577	P29401
881	Thiamine	228577	P51854
882	Thiamine	228577	Q9BU02
883	Thiamine	228577	Q9H3S4
884	Thiamylal	246459	P14867
885	Thiethylperazine	298783	P14416
886	Thiethylperazine	298783	P28223
887	Thiethylperazine	298783	P35348
888	Thioguanine	226995	P00492
889	Thioguanine	226995	P20839
890	Thioguanine	226995	Q06203
891	Thiopental	42425	Q693P7
892	Thioridazine	248121	P14416
893	Thioridazine	248121	P21728
894	Thioridazine	248121	P28223
895	Thioridazine	248121	P35348
896	Ticlopidine	247263	Q9H244
897	Timolol	246889	P07550
898	Timolol	246889	P08588
899	Tioconazole	121260	P10614
900	Tizanidine	162088	P08913
901	Tobramycin	235674	P0A7S3
902	Tocainide	246120	Q14524
903	Tolazamide	248643	P48048
904	Tolazoline	245997	P35348
905	Tolbutamide	190858	P48048
906	Tolbutamide	190858	Q09428
907	Tolmetin	80024	P23319
908	Tolmetin	80024	P35354
909	Topotecan	131707	Q969P6
910	Tramadol	158431	P23975
911	Tramadol	158431	P35372
912	Tranexamic Acid	229986	P00747
913	Tranylcypromine	3091	P21397
914	Trazodone	248796	P08913
915	Trazodone	248796	P23975
916	Trazodone	248796	P28223
917	Trazodone	248796	P31645
918	Trazodone	248796	P35348
919	Trazodone	248796	P35367
920	Tretinoin	51834	P10276
921	Tretinoin	51834	P10826
922	Tretinoin	51834	P13631
923	Triamcinolone	291995	P04083
924	Triamcinolone	291995	P08185
925	Triamterene	241809	Q13621
926	Triazolam	79742	P30536
927	Trichlormethiazide	234355	P00915
928	Trichlormethiazide	234355	P00918
929	Trichlormethiazide	234355	P05023
930	Trichlormethiazide	234355	P22748

continued on next page.

continued from previous page.

	chemical compound	NIST ID of compound	UniProt ID of protein
931	Trichlormethiazide	234355	Q12791
932	Trichlormethiazide	234355	Q13621
933	Trifluoperazine	17112	P14416
934	Trifluoperazine	17112	P35348
935	Triflupromazine	113868	P08172
936	Triflupromazine	113868	P11229
937	Triflupromazine	113868	P14416
938	Triflupromazine	113868	P21728
939	Triflupromazine	113868	P41595
940	Trihexyphenidyl	244993	P11229
941	Trilostane	298789	P14060
942	Trimeprazine	42348	P35367
943	Trimethadione	113821	O43497
944	Trimethoprim	247707	P08192
945	Trimethoprim	247707	Q27713
946	Trimipramine	235686	P23975
947	Trimipramine	235686	P31645
948	Tripelennamine	250622	P35367
949	Triprolidine	291072	P35367
950	Tropicamide	248908	P08173
951	Tubocurarine	234187	Q15822
952	Valproic Acid	191428	P80404
953	Venlafaxine	247381	P23975
954	Venlafaxine	247381	P31645
955	Verapamil	247152	P00915
956	Verapamil	247152	Q14524
957	Vitamin A	238604	O94788
958	Vitamin A	238604	P00352
959	Vitamin A	238604	P02753
960	Vitamin A	238604	P09455
961	Vitamin A	238604	P10745
962	Vitamin A	238604	P12271
963	Vitamin A	238604	P50120
964	Vitamin A	238604	P82980
965	Vitamin A	238604	Q8NBN7
966	Vitamin A	238604	Q92781
967	Vitamin A	238604	Q96NR8
968	Vitamin A	238604	Q96R05
969	Vitamin C	228563	P13674
970	Vitamin C	228563	Q02809
971	Vitamin C	228563	Q9UGH3
972	Vitamin C	228563	Q9UHI7
973	Warfarin	191650	Q9BQB6
974	Zidovudine	248197	P03369
975	Zolpidem	247600	P14867
976	Zolpidem	247600	P30536
977	Zonisamide	247580	O43497
978	Zonisamide	247580	P00915
979	Zonisamide	247580	Q14524
980	Zopiclone	247516	P30536

Supplementary Table A39 Positive pairs in DrugBank2 dataset

	UniProt ID	PubChem ID
1	O00180	60753
2	O00555	2351
3	O00555	2761
4	O00555	3955
5	O00555	4499
6	O00555	5486971
7	O00764	1054
8	O14646	41867
9	O14659	60843
10	O14987	5362391
11	O15399	3955
12	O15399	4601
13	O15528	134070
14	O15528	5751
15	O15528	6433735
16	O15554	3562
17	O30585	4173
18	O43497	16362
19	O43497	3291
20	O43497	441341
21	O43497	5576
22	O43497	5734
23	O43772	10917
24	O60391	4601
25	O60779	1130
26	O76074	110634
27	O76074	110635
28	O76074	5212
29	O76082	10917
30	O90777	148192
31	O90777	392622
32	O90777	5362440
33	O90777	60787
34	O90777	64143
35	O90777	65016
36	O90777	65027
37	O94788	1071
38	O95069	71329
39	O95665	54385
40	O95820	4911
41	P00325	3406
42	P00326	3406
43	P00352	1071
44	P00363	5430
45	P00374	126941
46	P00374	5583
47	P00374	60843
48	P00390	2578
49	P00395	3052
50	P00439	1125
51	P00440	2266
52	P00491	20279
53	P00491	667490
54	P00492	2723601
55	P00492	667490
56	P00519	5291
57	P00533	123631
58	P00533	176870
59	P00582	2266
60	P00734	152951
61	P00734	772
62	P00742	772
63	P00747	5526
64	P00747	564
65	P00797	5462340
66	P00807	42617
67	P00813	40926
68	P00915	1986
69	P00915	2162
70	P00915	2315
71	P00915	2343
72	P00915	2520
73	P00915	2720
74	P00915	2910
75	P00915	3019
76	P00915	3038
77	P00915	3284
78	P00915	3295
79	P00915	3639
80	P00915	3647
81	P00915	4100
82	P00915	4121
83	P00915	441341
84	P00915	5560
85	P00915	5734
86	P00915	68844
87	P00918	2315
88	P00918	2343
89	P00918	2720
90	P00918	2910
91	P00918	3019
92	P00918	3154
93	P00918	3284
94	P00918	3639
95	P00918	3647
96	P00918	4121
97	P00918	5284627
98	P00918	5560
99	P00918	68844
100	P00956	6476007
101	P01008	636380
102	P01008	772
103	P01130	157922
104	P01130	57166
105	P01178	439302
106	P01375	5426
107	P02753	1071
108	P02766	5819
109	P02766	5920
110	P02768	16850
111	P02768	5920
112	P02918	150610
113	P02919	150610
114	P03200	12620
115	P03369	18283
116	P03369	24066
117	P03369	35370
118	P03369	4463
119	P03369	464205
120	P03369	50599
121	P03369	5625
122	P03369	60825
123	P03369	60877
124	P03369	64139
125	P03369	65140
126	P03372	104741
127	P03372	10631
128	P03372	11289
129	P03372	13109
130	P03372	1548955
131	P03372	18140
132	P03372	19090
133	P03372	224004
134	P03372	3005573
135	P03372	3049
136	P03372	3054
137	P03372	40973
138	P03372	40976
139	P03372	5035
140	P03372	5360970
141	P03372	5376
142	P03372	5757
143	P03372	5870
144	P03372	5991
145	P03372	5994
146	P03372	6446
147	P03372	9270
148	P03372	9919
149	P03886	3226
150	P03886	3562
151	P03886	3763
152	P03886	4116
153	P03886	42113
154	P03886	5206
155	P04035	1548972
156	P04035	446156
157	P04035	53232
158	P04035	54454
159	P04035	54687
160	P04035	60823
161	P04035	6439133
162	P04049	216239
163	P04070	5248
164	P04083	123620
165	P04083	20469
166	P04083	31307
167	P04083	32798
168	P04083	39507
169	P04083	408334
170	P04083	443958
171	P04083	443980
172	P04083	444025
173	P04083	48175
174	P04083	52421
175	P04083	5282493
176	P04083	5311067
177	P04083	5743
178	P04083	5754
179	P04083	5865
180	P04083	636374

continued on next page.

continued from previous page.

	UniProt ID	PubChem ID
181	P04083	6741
182	P04083	71414
183	P04083	9782
184	P04150	123620
185	P04150	408334
186	P04150	443958
187	P04150	444025
188	P04150	5743
189	P04150	5754
190	P04150	5865
191	P04150	63006
192	P04150	6741
193	P04150	9782
194	P04278	224004
195	P04278	6446
196	P04293	2022
197	P04293	3324
198	P04293	3415
199	P04293	3454
200	P04293	4725
201	P04293	5905
202	P04293	60773
203	P04293	6256
204	P04746	2329
205	P04746	441184
206	P04746	441314
207	P04818	104758
208	P04818	3385
209	P04818	54575
210	P04818	5790
211	P04818	60750
212	P04818	60843
213	P04818	60953
214	P05023	2471
215	P05023	30322
216	P05023	3440
217	P05023	3647
218	P05023	4121
219	P05023	439501
220	P05023	443932
221	P05023	4679
222	P05023	5560
223	P05023	68949
224	P05041	5320
225	P05091	3117
226	P05141	25419
227	P05230	37720
228	P05543	5819
229	P05543	5920
230	P06149	3598
231	P06276	10547
232	P06276	5353894
233	P06276	5936
234	P06276	5965
235	P06276	656508
236	P06276	77991
237	P06276	9434
238	P06401	10631
239	P06401	13109
240	P06401	19090
241	P06401	40973
242	P06401	40976
243	P06401	5360970
244	P06401	55245
245	P06401	5994
246	P06401	6230
247	P06401	9051
248	P06401	9270
249	P06478	4725
250	P06715	1839
251	P06746	6253
252	P06818	60855
253	P06858	2796
254	P06858	3463
255	P06864	11333
256	P06982	287180
257	P07003	1839
258	P07101	1125
259	P07101	441350
260	P07140	4004
261	P07202	1349907
262	P07202	31072
263	P07202	657298
264	P07327	3406
265	P07437	13342
266	P07437	5978
267	P07437	60780
268	P07550	2083
269	P07550	2119
270	P07550	2405
271	P07550	2431
272	P07550	2583
273	P07550	2995
274	P07550	31477
275	P07550	33572
276	P07550	33624
277	P07550	3410
278	P07550	35330
279	P07550	3779
280	P07550	3869
281	P07550	39147
282	P07550	39468
283	P07550	4086
284	P07550	4828
285	P07550	4946
286	P07550	5152
287	P07550	5403
288	P07550	5816
289	P07550	7028
290	P08172	174174
291	P08172	20299
292	P08172	2342
293	P08172	2551
294	P08172	2995
295	P08172	3055
296	P08172	3354
297	P08172	4167
298	P08172	43232
299	P08172	441290
300	P08172	4585
301	P08172	4642
302	P08172	4919
303	P08172	4926
304	P08172	4927
305	P08172	4940
306	P08172	5440
307	P08172	5568
308	P08172	60168
309	P08172	60774
310	P08172	64692
311	P08172	6646
312	P08172	6647
313	P08172	6726
314	P08172	67425
315	P08172	71316
316	P08173	174174
317	P08173	2342
318	P08173	4167
319	P08173	4585
320	P08173	4919
321	P08173	4926
322	P08173	4927
323	P08173	4940
324	P08173	5440
325	P08173	5593
326	P08173	6646
327	P08185	15209
328	P08185	20469
329	P08185	247839
330	P08185	31307
331	P08185	39507
332	P08185	443980
333	P08185	444036
334	P08185	48175
335	P08185	5755
336	P08185	6215
337	P08185	636374
338	P08185	82153
339	P08185	9642
340	P08185	9878
341	P08192	5329
342	P08192	5578
343	P08235	13126
344	P08235	150310
345	P08235	31378
346	P08253	119031
347	P08473	126046
348	P08473	5362417
349	P08514	60947
350	P08546	3415
351	P08546	60613
352	P08588	1978
353	P08588	2119
354	P08588	2157
355	P08588	2249
356	P08588	2369
357	P08588	2405
358	P08588	2431
359	P08588	2583
360	P08588	2585

continued on next page.

continued from previous page.

	UniProt ID	PubChem ID
361	P08588	2995
362	P08588	31477
363	P08588	33624
364	P08588	36811
365	P08588	3762
366	P08588	3779
367	P08588	3869
368	P08588	39147
369	P08588	39468
370	P08588	4171
371	P08588	4828
372	P08588	4946
373	P08588	5073
374	P08588	5816
375	P08588	59768
376	P08588	60789
377	P08588	60854
378	P08588	681
379	P08620	37720
380	P08700	2161
381	P08908	2477
382	P08908	5487301
383	P08908	592735
384	P08908	60795
385	P08908	60854
386	P08908	77993
387	P08912	174174
388	P08912	2342
389	P08912	4167
390	P08912	4585
391	P08912	4926
392	P08912	4927
393	P08912	4940
394	P08912	5440
395	P08912	6646
396	P08913	2216
397	P08913	2341
398	P08913	2368
399	P08913	2435
400	P08913	26934
401	P08913	2803
402	P08913	3105
403	P08913	3241
404	P08913	3341
405	P08913	3446
406	P08913	3517
407	P08913	3518
408	P08913	3519
409	P08913	38521
410	P08913	38853
411	P08913	4205
412	P08913	4636
413	P08913	5487
414	P08913	5533
415	P08913	5775
416	P08913	5826
417	P08913	68602
418	P08913	7028
419	P09172	681
420	P09252	2022
421	P09252	3324
422	P09455	1071
423	P09619	216239
424	P09619	5291
425	P09884	119182
426	P09884	30751
427	P09917	3052
428	P0A3M5	36272
429	P0A560	3279
430	P0A5Y6	2761171
431	P0A5Y6	3767
432	P0A674	5381226
433	P0A674	5462354
434	P0A6B4	401
435	P0A6J8	401
436	P0A6R0	28517
437	P0A749	446987
438	P0A7J6	29029
439	P0A7J6	298
440	P0A7J6	3032400
441	P0A7J6	444037
442	P0A7J6	5281011
443	P0A7J6	5281054
444	P0A7J6	55185
445	P0A7J6	84029
446	P0A7R5	5353830
447	P0A7S3	15541
448	P0A7S3	19649
449	P0A7S3	3467
450	P0A7S3	441188
451	P0A7S3	441306
452	P0A7S3	5496
453	P0A7S3	6032
454	P0A7S3	8378
455	P0A7S3	9346
456	P0A7V8	5280963
457	P0A7V8	5280972
458	P0A7V8	5281011
459	P0A7V8	5281021
460	P0A7V8	5282044
461	P0A7V8	5311063
462	P0A7V8	5353990
463	P0A7V8	5464321
464	P0A7X3	5280963
465	P0A7X3	5280972
466	P0A7X3	5281011
467	P0A7X3	5281021
468	P0A7X3	5282044
469	P0A7X3	5311063
470	P0A7X3	5353990
471	P0A7X3	5464321
472	P0A7Z4	6323490
473	P0A8V2	5381226
474	P0A8V2	6436173
475	P10109	4211
476	P10275	224004
477	P10275	229455
478	P10275	3397
479	P10275	56069
480	P10275	5878
481	P10275	6013
482	P10275	6446
483	P10276	3312
484	P10276	444795
485	P10276	5381
486	P10276	5538
487	P10276	60164
488	P10276	6437841
489	P10415	148124
490	P10415	36314
491	P10613	441086
492	P10614	2812
493	P10614	3198
494	P10614	3365
495	P10614	441383
496	P10614	47576
497	P10614	5482
498	P10614	65863
499	P10614	71616
500	P10721	216239
501	P10721	5291
502	P10745	1071
503	P10826	3312
504	P10826	444795
505	P10826	5381
506	P10826	5538
507	P10826	60164
508	P10826	6437841
509	P10827	5819
510	P10827	5920
511	P10828	5920
512	P10875	441140
513	P11229	107979
514	P11229	174174
515	P11229	20299
516	P11229	2342
517	P11229	2370
518	P11229	2381
519	P11229	2551
520	P11229	2564
521	P11229	2784
522	P11229	2905
523	P11229	2911
524	P11229	2995
525	P11229	3042
526	P11229	3055
527	P11229	3162
528	P11229	3290
529	P11229	3354
530	P11229	3494
531	P11229	4097
532	P11229	4167
533	P11229	4168
534	P11229	43232
535	P11229	441342
536	P11229	443937
537	P11229	4585
538	P11229	4634
539	P11229	4642
540	P11229	4848

continued on next page.

continued from previous page.

	UniProt ID	PubChem ID
541	P11229	4919
542	P11229	4926
543	P11229	4927
544	P11229	4934
545	P11229	4940
546	P11229	5184
547	P11229	5314
548	P11229	5440
549	P11229	5568
550	P11229	5572
551	P11229	5749
552	P11229	5910
553	P11229	60774
554	P11229	64692
555	P11229	6646
556	P11229	6647
557	P11229	6726
558	P11229	6729
559	P11229	6832
560	P11229	83898
561	P11388	2179
562	P11388	31703
563	P11388	34698
564	P11388	36462
565	P11388	41744
566	P11388	41867
567	P11388	4212
568	P11388	42890
569	P11388	71384
570	P11473	134070
571	P11473	5281010
572	P11473	5751
573	P11473	6221
574	P11473	6433735
575	P11473	77996
576	P11485	65028
577	P11511	13769
578	P11511	2145
579	P11511	2187
580	P11511	3902
581	P11511	60198
582	P12235	25419
583	P12236	25419
584	P12268	4271
585	P12268	446541
586	P12271	1071
587	P12314	157922
588	P12314	57166
589	P12319	119212
590	P12461	3366
591	P12821	107807
592	P12821	32681
593	P12821	8550
594	P13631	3312
595	P13631	444795
596	P13631	5381
597	P13631	5538
598	P13631	60164
599	P13631	6437841
600	P13674	5785
601	P13716	137
602	P14060	656583
603	P14324	2088
604	P14324	4674
605	P14324	5245
606	P14324	60852
607	P14324	68740
608	P14416	16362
609	P14416	17012
610	P14416	2130
611	P14416	2477
612	P14416	2726
613	P14416	2729
614	P14416	2761
615	P14416	2818
616	P14416	31101
617	P14416	3151
618	P14416	3168
619	P14416	3372
620	P14416	3559
621	P14416	3964
622	P14416	4078
623	P14416	4168
624	P14416	4199
625	P14416	441185
626	P14416	4585
627	P14416	4748
628	P14416	47811
629	P14416	4917
630	P14416	4926
631	P14416	4927
632	P14416	4940
633	P14416	5002
634	P14416	5073
635	P14416	5095
636	P14416	5355
637	P14416	5440
638	P14416	54477
639	P14416	5452
640	P14416	54746
641	P14416	5566
642	P14416	5568
643	P14416	59868
644	P14416	6005
645	P14416	6047
646	P14416	60795
647	P14416	60854
648	P14780	119031
649	P14867	2118
650	P14867	2479
651	P14867	2481
652	P14867	2712
653	P14867	2789
654	P14867	2802
655	P14867	2809
656	P14867	3000715
657	P14867	3016
658	P14867	3032285
659	P14867	31143
660	P14867	31640
661	P14867	3226
662	P14867	3281
663	P14867	3373
664	P14867	3393
665	P14867	3562
666	P14867	36339
667	P14867	3763
668	P14867	4099
669	P14867	4116
670	P14867	4162
671	P14867	4192
672	P14867	42113
673	P14867	4616
674	P14867	4737
675	P14867	4763
676	P14867	4909
677	P14867	4943
678	P14867	5206
679	P14867	5284627
680	P14867	5360688
681	P14867	5361323
682	P14867	5719
683	P14867	5732
684	P14867	606974
685	P14867	702
686	P14867	71158
687	P14867	8271
688	P14867	9034
689	P14926	28517
690	P15056	216239
691	P15382	3702
692	P15538	4211
693	P16050	2749
694	P16050	60490
695	P16050	71398
696	P16066	145068
697	P16066	27661
698	P16066	4510
699	P16066	4512
700	P16066	6883
701	P16233	2329
702	P16233	3034010
703	P16471	224004
704	P16471	6446
705	P17117	5353830
706	P17752	1125
707	P18031	25419
708	P18031	60937
709	P18089	2368
710	P18089	3341
711	P18089	3518
712	P18089	38521
713	P18096	5362440
714	P18314	1990
715	P18405	152945
716	P18405	57363
717	P18825	2368
718	P18825	3341
719	P18825	3518
720	P18825	38521

continued on next page.

continued from previous page.

	UniProt ID	PubChem ID
721	P18858	456190
722	P18956	5361919
723	P19429	3033825
724	P19634	16231
725	P19793	3312
726	P19793	444795
727	P19793	60164
728	P19793	6437841
729	P19801	16231
730	P20061	441416
731	P20061	5479203
732	P20062	441416
733	P20062	5479203
734	P20082	287180
735	P20309	174174
736	P20309	20299
737	P20309	2342
738	P20309	3055
739	P20309	4167
740	P20309	444031
741	P20309	4585
742	P20309	4642
743	P20309	4926
744	P20309	4927
745	P20309	4940
746	P20309	5440
747	P20309	6126
748	P20309	6646
749	P20309	6647
750	P20309	6726
751	P20309	83898
752	P20648	3883
753	P20648	4594
754	P20711	34359
755	P20839	2723601
756	P20839	37542
757	P20839	4271
758	P20839	446541
759	P21233	12560
760	P21397	3675
761	P21397	3759
762	P21397	4235
763	P21397	441233
764	P21397	441401
765	P21397	4762
766	P21430	2130
767	P21430	5071
768	P21554	5284592
769	P21728	17012
770	P21728	2130
771	P21728	2729
772	P21728	28864
773	P21728	3341
774	P21728	4199
775	P21728	441185
776	P21728	4585
777	P21728	4748
778	P21728	47811
779	P21728	4926
780	P21728	4940
781	P21728	5440
782	P21728	5452
783	P21728	5568
784	P21728	5760
785	P21728	6005
786	P21728	6047
787	P21728	681
788	P21728	8226
789	P21731	5362391
790	P21817	2952
791	P21917	2818
792	P21917	4585
793	P21917	4926
794	P21917	4940
795	P21917	5440
796	P21917	6005
797	P21917	60854
798	P21918	3341
799	P21964	4659569
800	P21964	5281081
801	P22033	441416
802	P22303	1935
803	P22303	2131
804	P22303	3152
805	P22303	3202
806	P22303	4991
807	P22303	5965
808	P22303	9651
809	P22413	2141
810	P22413	37542
811	P22748	2315
812	P22748	2343
813	P22748	2720
814	P22748	2910
815	P22748	3019
816	P22748	3639
817	P22748	3647
818	P22748	4121
819	P22748	5284627
820	P22748	5560
821	P22888	47725
822	P22932	3312
823	P22932	444795
824	P22932	5381
825	P22932	60164
826	P22932	6437841
827	P22966	126046
828	P22966	44093
829	P22966	5362032
830	P22966	5362119
831	P22966	5362124
832	P22966	5362129
833	P22966	5484727
834	P22966	54892
835	P22966	55891
836	P22966	91270
837	P23219	1302
838	P23219	1983
839	P23219	2244
840	P23219	22881
841	P23219	2749
842	P23219	3033
843	P23219	3059
844	P23219	3308
845	P23219	3342
846	P23219	3394
847	P23219	3672
848	P23219	3715
849	P23219	3825
850	P23219	3826
851	P23219	4044
852	P23219	4075
853	P23219	4409
854	P23219	5280452
855	P23219	5281106
856	P23219	5312154
857	P23219	5359
858	P23219	5362070
859	P23219	5509
860	P23219	60726
861	P23219	71398
862	P23415	3226
863	P23415	3562
864	P23415	3763
865	P23415	4116
866	P23415	42113
867	P23415	5206
868	P23415	702
869	P23416	702
870	P23786	10917
871	P23786	4746
872	P23921	119182
873	P23921	30751
874	P23921	3657
875	P23921	60750
876	P23975	2160
877	P23975	2170
878	P23975	2368
879	P23975	2801
880	P23975	2995
881	P23975	33741
882	P23975	3518
883	P23975	3696
884	P23975	38521
885	P23975	4011
886	P23975	4020
887	P23975	4158
888	P23975	444
889	P23975	4449
890	P23975	4543
891	P23975	4762
892	P23975	4771
893	P23975	4976
894	P23975	5210
895	P23975	54841
896	P23975	5533
897	P23975	5584
898	P23975	5656
899	P23975	5760
900	P23975	60835

continued on next page.

continued from previous page.

	UniProt ID	PubChem ID
901	P23975	65856
902	P23975	667477
903	P23975	7029
904	P24024	153941
905	P24024	60871
906	P24046	5360688
907	P24228	150610
908	P24228	64778
909	P24530	104865
910	P24752	5353980
911	P25021	2756
912	P25021	3001055
913	P25021	3033637
914	P25021	3241
915	P25021	3325
916	P25021	65513
917	P25021	7741
918	P25100	129211
919	P25100	2368
920	P25100	3033538
921	P25100	3518
922	P25100	38521
923	P25100	4926
924	P25100	4940
925	P25101	104865
926	P25103	151165
927	P26282	5328
928	P26676	37542
929	P27338	26757
930	P27338	3759
931	P27707	60750
932	P28221	10531
933	P28221	123606
934	P28221	441240
935	P28221	4440
936	P28221	5078
937	P28221	5358
938	P28221	5487301
939	P28221	60854
940	P28221	77992
941	P28221	77993
942	P28222	10531
943	P28222	441240
944	P28222	4440
945	P28222	5078
946	P28222	5358
947	P28222	77992
948	P28222	77993
949	P28222	8223
950	P28223	2726
951	P28223	2818
952	P28223	2895
953	P28223	2913
954	P28223	3241
955	P28223	3964
956	P28223	4078
957	P28223	4199
958	P28223	4205
959	P28223	4449
960	P28223	4585
961	P28223	4917
962	P28223	4926
963	P28223	4927
964	P28223	4940
965	P28223	5002
966	P28223	5073
967	P28223	5440
968	P28223	5452
969	P28223	5533
970	P28223	60795
971	P28223	60854
972	P28223	9681
973	P28335	3337
974	P28335	4199
975	P28335	4205
976	P28335	4585
977	P28335	4926
978	P28335	4940
979	P28335	5002
980	P28335	5440
981	P28335	60854
982	P28566	77993
983	P28702	3312
984	P28702	444795
985	P28702	5381
986	P28702	60164
987	P28702	6437841
988	P28702	82146
989	P29251	2955
990	P29275	1676
991	P29275	2153
992	P29401	1130
993	P29466	5852
994	P29474	1125
995	P30043	6759
996	P30049	3226
997	P30049	3562
998	P30049	3763
999	P30049	4116
1000	P30049	42113
1001	P30049	5206
1002	P30085	60750
1003	P30273	119212
1004	P30518	151171
1005	P30518	24774
1006	P30518	27991
1007	P30531	5760
1008	P30531	60648
1009	P30536	2118
1010	P30536	2717
1011	P30536	2809
1012	P30536	3016
1013	P30536	3117
1014	P30536	31640
1015	P30536	3261
1016	P30536	37632
1017	P30536	3958
1018	P30536	5391
1019	P30536	5556
1020	P30536	5719
1021	P30536	5732
1022	P30536	5735
1023	P30536	969472
1024	P30542	2519
1025	P30542	3446
1026	P30542	3696
1027	P30556	130881
1028	P30556	2541
1029	P30556	3749
1030	P30556	3961
1031	P30556	5833
1032	P30556	60846
1033	P30556	60879
1034	P30556	65999
1035	P30939	77993
1036	P30968	36523
1037	P30968	3911
1038	P31213	2266
1039	P31644	2118
1040	P31645	146570
1041	P31645	2160
1042	P31645	2170
1043	P31645	2771
1044	P31645	2801
1045	P31645	2995
1046	P31645	3337
1047	P31645	3386
1048	P31645	3404
1049	P31645	3696
1050	P31645	4199
1051	P31645	43815
1052	P31645	4449
1053	P31645	4543
1054	P31645	4976
1055	P31645	5210
1056	P31645	5360696
1057	P31645	5533
1058	P31645	5584
1059	P31645	5656
1060	P31645	5760
1061	P31645	60835
1062	P31645	667477
1063	P31645	68617
1064	P32238	168399
1065	P32754	115355
1066	P33527	5342
1067	P33765	9433
1068	P34896	1054
1069	P34896	440473
1070	P34903	2118
1071	P34969	60854
1072	P34969	77993
1073	P34972	5284592
1074	P34995	149351
1075	P34995	2474
1076	P34995	5284525
1077	P35348	129211
1078	P35348	13109
1079	P35348	17012
1080	P35348	2092

continued on next page.

continued from previous page.

	UniProt ID	PubChem ID
1081	P35348	2157
1082	P35348	2341
1083	P35348	2368
1084	P35348	2585
1085	P35348	26934
1086	P35348	3033538
1087	P35348	3157
1088	P35348	3241
1089	P35348	34040
1090	P35348	3518
1091	P35348	38521
1092	P35348	3869
1093	P35348	4011
1094	P35348	4195
1095	P35348	4449
1096	P35348	4493
1097	P35348	4636
1098	P35348	4748
1099	P35348	4768
1100	P35348	4893
1101	P35348	4926
1102	P35348	4927
1103	P35348	4940
1104	P35348	5073
1105	P35348	5401
1106	P35348	5452
1107	P35348	5504
1108	P35348	5533
1109	P35348	5566
1110	P35348	5816
1111	P35348	5826
1112	P35348	5906
1113	P35348	6041
1114	P35348	6082
1115	P35348	60854
1116	P35348	7028
1117	P35348	8223
1118	P35354	119607
1119	P35354	1302
1120	P35354	1983
1121	P35354	216326
1122	P35354	2244
1123	P35354	22881
1124	P35354	2581
1125	P35354	2662
1126	P35354	2749
1127	P35354	3033
1128	P35354	3059
1129	P35354	3308
1130	P35354	3342
1131	P35354	3394
1132	P35354	3672
1133	P35354	3715
1134	P35354	3825
1135	P35354	3826
1136	P35354	4044
1137	P35354	4075
1138	P35354	4409
1139	P35354	4614
1140	P35354	4781
1141	P35354	5090
1142	P35354	5281106
1143	P35354	5312154
1144	P35354	5352
1145	P35354	5359
1146	P35354	5362070
1147	P35354	5509
1148	P35354	60726
1149	P35354	71398
1150	P35367	124087
1151	P35367	14677
1152	P35367	16960
1153	P35367	19861
1154	P35367	2247
1155	P35367	2267
1156	P35367	2342
1157	P35367	2444
1158	P35367	2564
1159	P35367	2678
1160	P35367	26987
1161	P35367	2725
1162	P35367	2761
1163	P35367	2818
1164	P35367	2995
1165	P35367	3100
1166	P35367	3103
1167	P35367	3162
1168	P35367	3219
1169	P35367	3241
1170	P35367	3348
1171	P35367	3658
1172	P35367	3827
1173	P35367	3957
1174	P35367	4011
1175	P35367	4034
1176	P35367	4066
1177	P35367	441281
1178	P35367	4585
1179	P35367	4601
1180	P35367	4917
1181	P35367	4926
1182	P35367	4927
1183	P35367	4940
1184	P35367	50294
1185	P35367	5073
1186	P35367	5281071
1187	P35367	5282443
1188	P35367	5405
1189	P35367	54385
1190	P35367	5440
1191	P35367	5533
1192	P35367	5574
1193	P35367	5587
1194	P35367	57697
1195	P35367	60854
1196	P35367	65513
1197	P35367	667477
1198	P35367	6726
1199	P35367	6729
1200	P35367	6834
1201	P35368	129211
1202	P35368	2368
1203	P35368	3033538
1204	P35368	3518
1205	P35368	38521
1206	P35368	3869
1207	P35368	4236
1208	P35368	4926
1209	P35368	4940
1210	P35368	5401
1211	P35372	13505
1212	P35372	15130
1213	P35372	15330
1214	P35372	3345
1215	P35372	33741
1216	P35372	3955
1217	P35372	40400
1218	P35372	40841
1219	P35372	4095
1220	P35372	41693
1221	P35372	441278
1222	P35372	51263
1223	P35372	5284371
1224	P35372	5284569
1225	P35372	5284570
1226	P35372	5284596
1227	P35372	5284603
1228	P35372	5284604
1229	P35372	5288826
1230	P35372	5359272
1231	P35372	5359371
1232	P35372	5360515
1233	P35372	5360630
1234	P35372	5361092
1235	P35372	60815
1236	P35372	8944
1237	P35462	5095
1238	P35462	5760
1239	P35462	59868
1240	P35462	60854
1241	P35498	441341
1242	P35498	4753
1243	P35498	5284627
1244	P35498	5734
1245	P35520	1054
1246	P35610	72281
1247	P35916	216239
1248	P35968	216239
1249	P36888	216239
1250	P37088	16231
1251	P37231	4829
1252	P37231	5591
1253	P37231	77999
1254	P37243	5920
1255	P37254	5333
1256	P37288	151171
1257	P37288	24774
1258	P37288	5956
1259	P38435	2197
1260	P38435	4812

continued on next page.

continued from previous page.

	UniProt ID	PubChem ID
1261	P39086	5284627
1262	P39435	28517
1263	P40030	14956
1264	P40030	441142
1265	P40030	441382
1266	P40030	5361463
1267	P40261	938
1268	P41143	15330
1269	P41143	16362
1270	P41143	5284371
1271	P41143	5284569
1272	P41143	5284570
1273	P41143	5284603
1274	P41143	5360630
1275	P41145	15330
1276	P41145	40400
1277	P41145	4058
1278	P41145	40841
1279	P41145	5284371
1280	P41145	5284569
1281	P41145	5284570
1282	P41145	5284603
1283	P41145	5360630
1284	P41145	5361092
1285	P41145	5760
1286	P41180	156419
1287	P41595	4199
1288	P41595	5002
1289	P41595	5568
1290	P41595	77993
1291	P42261	3226
1292	P42261	3562
1293	P42261	3763
1294	P42261	4116
1295	P42261	42113
1296	P42261	5206
1297	P42345	6436030
1298	P42971	456256
1299	P42971	54116
1300	P43088	60798
1301	P43088	656627
1302	P43116	149351
1303	P43116	6443959
1304	P43119	5282381
1305	P43119	5282415
1306	P43119	6443959
1307	P43119	9691
1308	P43155	10917
1309	P43700	149096
1310	P43700	152946
1311	P43700	2764
1312	P43700	3229
1313	P43700	3948
1314	P43700	4539
1315	P43700	4583
1316	P43700	51081
1317	P43700	5379
1318	P43700	5464436
1319	P43700	60464
1320	P43700	62959
1321	P43700	72474
1322	P43702	149096
1323	P43702	152946
1324	P43702	2764
1325	P43702	3229
1326	P43702	3948
1327	P43702	4539
1328	P43702	4583
1329	P43702	51081
1330	P43702	5379
1331	P43702	5464436
1332	P43702	60464
1333	P43702	62959
1334	P43702	72474
1335	P44466	150610
1336	P44469	150610
1337	P45059	150610
1338	P45379	3033825
1339	P46098	148211
1340	P46098	2099
1341	P46098	3510
1342	P46098	4205
1343	P46098	4595
1344	P46098	4914
1345	P46098	60654
1346	P46098	8612
1347	P47712	15209
1348	P47712	247839
1349	P47712	444036
1350	P47712	6215
1351	P47712	82153
1352	P47712	9642
1353	P47712	9878
1354	P47869	2118
1355	P47901	24774
1356	P47989	2094
1357	P48039	208902
1358	P48039	896
1359	P48048	1989
1360	P48048	2727
1361	P48048	3475
1362	P48048	3476
1363	P48048	3478
1364	P48048	3488
1365	P48048	4201
1366	P48048	5503
1367	P48048	5505
1368	P48048	60026
1369	P48048	65981
1370	P48169	2118
1371	P48443	3312
1372	P48443	444795
1373	P48443	60164
1374	P48443	6437841
1375	P49069	6435893
1376	P49286	208902
1377	P49902	37542
1378	P49916	456190
1379	P50120	1071
1380	P50416	10917
1381	P50416	4746
1382	P50859	4189
1383	P50859	47472
1384	P50859	55283
1385	P51164	5029
1386	P51168	16231
1387	P51170	16231
1388	P51172	16231
1389	P51649	2716
1390	P51787	3702
1391	P51788	656719
1392	P51854	1130
1393	P54289	24083
1394	P54289	3333
1395	P54289	3784
1396	P54289	4485
1397	P54289	60753
1398	P54710	2910
1399	P55011	24015
1400	P55017	2315
1401	P55017	2343
1402	P55017	2720
1403	P55017	3019
1404	P55017	3639
1405	P55157	72281
1406	P55263	37542
1407	P55773	54786
1408	P58596	439542
1409	P61889	1839
1410	P62158	3955
1411	P62942	439492
1412	P62942	6436030
1413	P63098	6447131
1414	P63252	3033825
1415	P64143	4649
1416	P66010	2266
1417	P68344	12620
1418	P69905	2165
1419	P69905	2719
1420	P69905	4046
1421	P69905	4908
1422	P69905	8549
1423	P72030	3279
1424	P72059	3279
1425	P77390	1839
1426	P78348	16231
1427	P80404	3121
1428	P80404	53519
1429	P80404	5665
1430	P82980	1071
1431	P87066	441140
1432	P98194	3226
1433	P98194	3562
1434	P98194	3763
1435	P98194	4116
1436	P98194	42113
1437	P98194	5206
1438	Q01650	3446
1439	Q01959	4020
1440	Q01959	4158

continued on next page.

continued from previous page.

	UniProt ID	PubChem ID
1441	Q01959	4236
1442	Q01959	444
1443	Q01959	4762
1444	Q01959	4771
1445	Q01959	4914
1446	Q01959	5760
1447	Q01959	60835
1448	Q01959	7029
1449	Q01959	8612
1450	Q02127	3899
1451	Q02318	6221
1452	Q02641	60753
1453	Q02809	5785
1454	Q05486	60825
1455	Q05586	4601
1456	Q05940	5770
1457	Q06203	2723601
1458	Q06432	2162
1459	Q06432	24083
1460	Q06432	39186
1461	Q06432	4497
1462	Q06432	4507
1463	Q06432	60753
1464	Q06432	65866
1465	Q07343	1676
1466	Q07343	2153
1467	Q07343	2519
1468	Q07343	2754
1469	Q07343	3182
1470	Q07343	4197
1471	Q07343	4680
1472	Q07343	4740
1473	Q07869	3339
1474	Q08129	2761171
1475	Q08129	3767
1476	Q08828	60961
1477	Q09428	3475
1478	Q09428	5505
1479	Q09470	3226
1480	Q09470	3763
1481	Q09470	4116
1482	Q09470	42113
1483	Q09470	5206
1484	Q11195	1046
1485	Q12791	2315
1486	Q12791	2343
1487	Q12791	2720
1488	Q12791	2733
1489	Q12791	27686
1490	Q12791	2910
1491	Q12791	3019
1492	Q12791	3639
1493	Q12791	3647
1494	Q12791	4121
1495	Q12791	5560
1496	Q12806	60947
1497	Q12809	4932
1498	Q12809	60753
1499	Q12879	3955
1500	Q13200	93860
1501	Q13224	3955
1502	Q13258	50294
1503	Q13258	5311027
1504	Q13621	24015
1505	Q13621	2471
1506	Q13621	2716
1507	Q13621	2732
1508	Q13621	3278
1509	Q13621	3440
1510	Q13621	3647
1511	Q13621	4121
1512	Q13621	4170
1513	Q13621	41781
1514	Q13621	5546
1515	Q13621	5560
1516	Q13639	2769
1517	Q13639	5487301
1518	Q13936	24083
1519	Q13936	4474
1520	Q13936	60753
1521	Q13946	3827
1522	Q14181	5351166
1523	Q14432	2153
1524	Q14432	2182
1525	Q14432	3033825
1526	Q14432	9433
1527	Q14524	10770
1528	Q14524	1775
1529	Q14524	2351
1530	Q14524	2520
1531	Q14524	2554
1532	Q14524	3025
1533	Q14524	3114
1534	Q14524	3292
1535	Q14524	3356
1536	Q14524	34312
1537	Q14524	3446
1538	Q14524	34633
1539	Q14524	3676
1540	Q14524	3878
1541	Q14524	38945
1542	Q14524	4060
1543	Q14524	4178
1544	Q14524	441074
1545	Q14524	48033
1546	Q14524	4906
1547	Q14524	4913
1548	Q14524	4932
1549	Q14524	5070
1550	Q14524	52195
1551	Q14524	5760
1552	Q14524	7699
1553	Q14534	2484
1554	Q14534	5402
1555	Q14534	73342
1556	Q14654	60753
1557	Q14894	5819
1558	Q14957	3955
1559	Q15274	938
1560	Q15822	2381
1561	Q15822	24244
1562	Q15822	2551
1563	Q15822	2968
1564	Q15822	4032
1565	Q15822	441290
1566	Q15822	47319
1567	Q15822	5359371
1568	Q15822	5360696
1569	Q15822	6000
1570	Q15822	60168
1571	Q15822	62886
1572	Q15822	67425
1573	Q15822	71316
1574	Q15822	942
1575	Q16445	2118
1576	Q16515	16231
1577	Q16647	4781
1578	Q16739	51634
1579	Q17328	6443957
1580	Q49184	5336
1581	Q5L2G3	2794
1582	Q5SQR9	3562
1583	Q5T1J1	5360696
1584	Q5TZY3	2801
1585	Q6R3Q4	150311
1586	Q6VVX0	5751
1587	Q6VVX0	6221
1588	Q75NA5	727
1589	Q76NM6	37393
1590	Q7TQM4	72281
1591	Q8ITG2	4837
1592	Q8ITG2	6443957
1593	Q8IVH4	441416
1594	Q8IVH4	5479203
1595	Q8N8R3	10917
1596	Q8NBN7	1071
1597	Q8TCU5	3290
1598	Q8TCU5	3331
1599	Q8TCU5	3446
1600	Q8TCU5	3821
1601	Q8TCU5	4054
1602	Q8TCU5	4058
1603	Q8TCU5	4095
1604	Q8TCU5	4601
1605	Q8TCU5	4914
1606	Q8TCU5	5070
1607	Q8TCU5	5360696
1608	Q8TCU5	702
1609	Q8TCU5	71158
1610	Q8TCU5	8612
1611	Q8TDS4	938
1612	Q8XJ01	107654
1613	Q8XJ01	18381
1614	Q8XJ01	19003
1615	Q8XJ01	19150
1616	Q8XJ01	20824
1617	Q8XJ01	21319
1618	Q8XJ01	2610
1619	Q8XJ01	27447
1620	Q8XJ01	30699

continued on next page.

continued from previous page.

	UniProt ID	PubChem ID
1621	Q8XJ01	33613
1622	Q8XJ01	37625
1623	Q8XJ01	3956
1624	Q8XJ01	40958
1625	Q8XJ01	42008
1626	Q8XJ01	43507
1627	Q8XJ01	43672
1628	Q8XJ01	43708
1629	Q8XJ01	443387
1630	Q8XJ01	444006
1631	Q8XJ01	456256
1632	Q8XJ01	5281006
1633	Q8XJ01	5361202
1634	Q8XJ01	5361486
1635	Q8XJ01	54362
1636	Q8XJ01	5904
1637	Q8XJ01	6024
1638	Q8XJ01	6098
1639	Q8XJ01	6249
1640	Q8XJ01	6399253
1641	Q8XJ01	656511
1642	Q8XJ01	6869
1643	Q8XJ01	8982
1644	Q8XJ01	91713
1645	Q92731	18140
1646	Q92781	1071
1647	Q968X7	41684
1648	Q969G6	6759
1649	Q969P6	60700
1650	Q969P6	60838
1651	Q96EY8	441416
1652	Q96EY8	5479203
1653	Q96GD0	1054
1654	Q96NR8	1071
1655	Q96P88	36523
1656	Q96R05	1071
1657	Q96S41	518740
1658	Q99460	93860
1659	Q99707	441416
1660	Q99707	5479203
1661	Q99720	5360696
1662	Q9BQB6	4760
1663	Q9BQB6	653
1664	Q9BQB6	6691
1665	Q9BQB6	9908
1666	Q9BU02	1130
1667	Q9BXJ7	441416
1668	Q9BXJ7	5479203
1669	Q9FAW5	5362070
1670	Q9GZZ6	23576
1671	Q9GZZ6	4095
1672	Q9GZZ6	4914
1673	Q9GZZ6	5850
1674	Q9GZZ6	8612
1675	Q9H015	10917
1676	Q9H244	114805
1677	Q9H244	5472
1678	Q9H244	54786
1679	Q9H244	60606
1680	Q9H252	60753
1681	Q9H2X9	24015
1682	Q9H3N8	2818
1683	Q9H3S4	1130
1684	Q9H4B7	148124
1685	Q9H4B7	36314
1686	Q9H4B7	40839
1687	Q9N587	26879
1688	Q9NP56	3827
1689	Q9NS40	60753
1690	Q9NS75	50294
1691	Q9NVS9	1054
1692	Q9NYK1	57469
1693	Q9NYX4	2818
1694	Q9P0X4	8280
1695	Q9QNF7	60773
1696	Q9QNF7	6256
1697	Q9UBK8	441416
1698	Q9UBK8	5479203
1699	Q9UBS5	2284
1700	Q9UGH3	5785
1701	Q9UHI7	5785
1702	Q9UHW9	24015
1703	Q9UI33	5760
1704	Q9UKG9	10917
1705	Q9UP95	24015
1706	Q9UPY5	5353980
1707	Q9Y233	3108
1708	Q9Y257	60753
1709	Q9Y271	2161
1710	Q9Y271	50294
1711	Q9Y271	5717
1712	Q9Y271	60951
1713	Q9Y271	6436135
1714	Q9Y478	4091
1715	Q9Y478	8249
1716	Q9Y5Y9	10770
1717	Q9Y5Y9	2337
1718	Q9Y5Y9	2474
1719	Q9Y5Y9	3025
1720	Q9Y5Y9	3180
1721	Q9Y5Y9	3676
1722	Q9Y5Y9	4062
1723	Q9Y5Y9	4633
1724	Q9Y5Y9	4914
1725	Q9Y5Y9	4935
1726	Q9Y5Y9	5760
1727	Q9Y5Y9	71273
1728	Q9Y5Y9	8612
1729	Q9Y5Y9	92253
1730	Q9Y617	1054
1731	Q9Y666	24015

Supplementary Table A40 199 clusters used for protein description

cluster #	# of the members	members in the cluster #
1	56	CAA, CAC, CAD, CAE, CAG, CAH, CAK, CAN, CAP, CAQ, CAR, CAS, CAT, CCD, CCE, CCG, CCH, CCK, CCN, CCP, CCQ, CCR, CCS, CCT, CDH, CDL, CDM, CDV, CDY, CEH, CEM, CET, CEY, CGH, CGM, CGY, CHH, CHN, CHP, CHQ, CHS, CHT, CKY, CLN, CMN, CMS, CNN, CNV, CNY, CPQ, CPR, CPT, CPY, CRY, CSY, CTT
2	14	CAI, CAL, CAV, CFV, CII, CIL, CIT, CIV, CLS, CLT, CLV, CTV, CVV, CVY
3	15	CFG, CFP, CFS, CGI, CGL, CGV, CGW, CIN, CIP, CIS, CLP, CPV, CPW, CSV, CTY
4	19	CFF, CFI, CFL, CFM, CFW, CFY, CIM, CIW, CIY, CLL, CLM, CLW, CLY, CMM, CMV, CMW, CVW, CWW, CWY
5	2	CDD, CDN
6	8	CCC, CCF, CCI, CCL, CCM, CCV, CCW, CCY
7	41	CAF, CAM, CAW, CAY, CDF, CDI, CDW, CEF, CEI, CEL, CEV, CEW, CFH, CFK, CFN, CFQ, CFR, CFT, CHI, CHL, CHM, CHV, CHW, CHY, CIQ, CKM, CKW, CLQ, CMP, CMQ, CMR, CMT, CMY, CNW, CQV, CQW, CQY, CRW, CSW, CTW, CYY
8	6	CIK, CIR, CKL, CKV, CLR, CRV
9	2	GMW, GWW
10	6	GCF, GCI, GCL, GCM, GCW, GMM
11	21	GAF, GFF, GFI, GFL, GFM, GFV, GFW, GFY, GHI, GIM, GIW, GIY, GLM, GLW, GLY, GMV, GMY, GVW, GVY, GWY, GYY
12	7	GAI, GII, GIL, GIV, GLL, GLV, GVV
13	17	GAD, GAE, GAG, GAK, GAN, GAQ, GAS, GDE, GEE, GEG, GEH, GEK, GEN, GEQ, GER, GES, GET
14	2	GAA, GAL
15	77	GAC, GAH, GAR, GAT, GAV, GAY, GCC, GCD, GCE, GCG, GCH, GCK, GCN, GCP, GCQ, GCR, GCS, GCT, GCV, GCY, GDI, GDL, GDM, GDV, GEV, GFG, GFK, GFN, GFQ, GFR, GFS, GFT, GGI, GGL, GGM, GGV, GGW, GHH, GHM, GHT, GHV, GHY, GIK, GIN, GIP, GIQ, GIR, GIS, GIT, GKL, GKM, GKV, GLN, GLP, GLQ, GLR, GLS, GLT, GMN, GMP, GMQ, GMR, GMS, GMT, GNV, GPV, GQV, GQY, GRV, GRW, GSV, GSW, GSY, GTT, GTV, GTW, GTY
16	21	GDP, GDY, GFP, GGP, GGY, GHP, GKP, GKW, GKY, GNP, GNY, GPP, GPQ, GPR, GPS, GPT, GPW, GPY, GRY, SGP, SPP
17	15	GAM, GAW, GDF, GDW, GEF, GEI, GEL, GEM, GEW, GEY, GFH, GHL, GHW, GNW, GQW
18	44	GAP, GDD, GDG, GDH, GDK, GDN, GDQ, GDR, GDS, GDT, GEP, GGG, GGH, GGK, GGN, GGQ, GGR, GGS, GGT, GHK, GHN, GHQ, GHR, GHS, GKK, GKN, GKQ, GKR, GKS, GKT, GNN, GNQ, GNR, GNS, GNT, GQQ, GQR, GQS, GQT, GRS, GRT, GSS, GST, SGG
19	6	WKR, WNR, WQR, WRR, WRS, WRT
20	2	WER, WKK
21	24	WDK, WDQ, WDR, WDS, WDT, WGG, WGK, WGN, WGQ, WGR, WGS, WGT, WHK, WKN, WKQ, WKS, WKT, WNQ, WNS, WNT, WQQ, WQS, WSS, WST
22	11	VCC, VCD, VCG, VCH, VCI, VCN, VCQ, VCR, VCS, VCT, VCV
23	16	VCE, VDD, VDH, VDM, VDN, VEH, VEM, VFH, VFM, VHH, VHM, VHN, VHQ, VMM, VMN, VMQ
24	19	ICC, ICT, ICV, ISV, ITV, IVV, VCF, VCM, VCW, VCY, VFI, VFV, VIM, VIY, VLL, VLY, VMV, VVW, VVY
25	22	ICG, ICS, ISS, IST, TII, TIL, TIV, VCP, VDP, VDY, VGP, VGW, VGY, VHP, VIP, VKP, VNN, VNP, VPQ, VPR, VPS, VPT
26	6	VIK, VIR, VKL, VKV, VLR, VRV
27	12	VAD, VAE, VAG, VAN, VDE, VDG, VEE, VEG, VEK, VEN, VEQ, VES
28	82	ITT, VAA, VAC, VAF, VAH, VAI, VAK, VAL, VAM, VAP, VAQ, VAR, VAS, VAT, VAV, VAW, VAY, VCK, VCL, VDF, VDI, VDL, VDV, VEF, VEI, VEL, VET, VEV, VEY, VFG, VFK, VFN, VFQ, VFR, VFS, VFT, VGH, VGI, VGL, VGM, VGV, VHI, VHL, VHR, VHS, VHT, VHV, VHY, VIN, VIQ, VIS, VIT, VKM, VKY, VLM, VLN, VLP, VLQ, VLS, VLT, VMP, VMR, VMS, VMT, VMY, VNV, VNY, VPV, VQT, VQV, VQY, VRT, VRW, VRY, VSV, VSW, VSY, VTT, VTV, VTW, VTY, VYY
29	28	VDK, VDQ, VDR, VDS, VDT, VER, VGG, VGK, VGN, VGQ, VGR, VGS, VGT, VHK, VKN, VKQ, VKS, VKT, VNQ, VNR, VNS, VNT, VQQ, VQR, VQS, VRS, VSS, VST
30	3	VKK, VKR, VRR
31	10	ECF, ECI, ECL, ECM, ECV, ECW, ECY, EIY, EMV, EVY
32	7	EDD, EDG, EDK, EDN, EGK, ENN, ENR
33	9	EFI, EFL, EFV, EII, EIL, EIV, ELL, ELV, EVV
34	16	EAD, EAE, EAG, EAK, EAN, EAS, EDE, EEE, EEG, EEH, EEK, EEN, EEQ, EER, EES, EET
35	10	EDP, EEP, EGP, EKP, ENP, EPP, EPQ, EPR, EPS, EPT
36	4	ECE, EDM, EHM, EMM
37	12	EFF, EFM, EFW, EFY, EIM, EIW, ELM, ELW, EMW, EVW, EWW, EWY
38	5	EIK, EIR, EKV, ELR, ERV
39	88	EAA, EAC, EAF, EAH, EAI, EAL, EAM, EAP, EAQ, EAR, EAT, EAV, EAW, EAY, ECP, EDF, EDI, EDL, EDV, EDW, EDY, EEF, EEI, EEL, EEM, EEV, EEW, EEY, EFG, EFH, EFK, EFN, EFP, EFQ, EFR, EFS, EFT, EGI, EGL, EGM, EGV, EGW, EGY, EHI, EHL, EHP, EHV, EHW, EHY, EIN, EIP, EIQ, EIS, EIT, EKL, EKM, EKW, EKY, ELN, ELP, ELQ, ELS, ELT, ELY, EMN, EMP, EMQ, EMS, EMT, EMY, ENV, ENW, ENY, EPV, EPW, EPY, EQV, EQW, EQY, ERW, ERY, ESV, ESW, ESY, ETV, ETW, ETY, EYY
40	2	FIR, FRV
41	7	FHR, FKR, FNR, FQR, FRR, FRS, FRT

continued on next page.

continued from previous page.

cluster #	# of the members	members in the cluster #
42	4	FCC, FCI, FCM, FCV
43	43	FCD, FCG, FCK, FDH, FDN, FDQ, FDR, FDS, FDT, FGG, FGH, FGN, FGQ, FGR, FGS, FGT, FHH, FHK, FHN, FHQ, FHS, FHT, FIS, FKK, FKN, FKQ, FKS, FKT, FMN, FMR, FNQ, FNS, FNT, FNV, FQQ, FQS, FQT, FSS, FST, FTT, MPP, MPW, MPY
44	6	FCH, FCN, FCQ, FCR, FCS, FCT
45	11	CDE, CDG, CEE, CEG, CEK, CEN, CEQ, CES, CGG, CGN, HCC
46	7	CDP, CEP, CGP, CKP, CNP, CPP, CPS
47	6	MDF, MDW, MEF, MEW, MMP, MNW
48	10	MAA, MAD, MAE, MAF, MAL, MAM, MAW, MDL, MEL, MLM
49	5	MFF, MFL, MFM, MFW, MLW
50	4	MHW, MMW, MWW, MWY
51	3	MDE, MEE, MEM
52	3	IKK, IKR, IRR
53	33	ICD, ICH, ICK, ICN, ICQ, ICR, IDH, IDN, IDR, IGH, IGN, IGR, IHH, IHK, IHN, IHQ, IHR, IHS, IHT, IKN, INN, INQ, INR, INS, INT, IQR, IRS, IRT, VDW, VFP, VHW, VNW, VQW
54	23	IAD, IAG, IAH, IAK, IAN, IAQ, IAR, IAS, ICE, IDE, IDM, IEG, IEH, IEK, IEM, IEN, IEQ, IER, IES, IET, LCC, VEP, VEW
55	17	IAF, IAI, IAL, IAV, IFT, IFV, III, IIL, IIT, IIV, ILL, ILT, ILV, VFF, VFL, VIW, VLW
56	19	IAP, IDD, IDG, IDK, IDQ, IDS, IDT, IGG, IGK, IGQ, IGS, IGT, IKQ, IKS, IKT, IQQ, IQS, IQT, VKW
57	40	FVV, IAM, IAW, ICF, ICI, ICL, ICM, ICW, ICY, IFF, IFH, IFI, IFL, IFM, IFW, IFY, IHI, IHL, IHM, IHV, IHW, IIM, IIW, IIY, ILM, ILW, ILY, IMM, IMV, IMW, IMY, ITW, IVW, IVY, IWW, IWY, IYY, VFW, VMW, VWW
58	72	FTV, IAA, IAC, IAT, IAY, ICP, IDF, IDI, IDL, IDV, IDW, IDY, IEF, IEI, IEL, IEV, IEW, IEY, IFG, IFK, IFN, IFP, IFQ, IFR, IFS, IGI, IGL, IGM, IGV, IGW, IGY, IHY, IIK, IIN, IIP, IIQ, IIR, IIS, IKL, IKM, IKV, IKW, IKY, ILN, ILP, ILQ, ILR, ILS, IMN, IMP, IMQ, IMR, IMS, IMT, INV, INW, INY, IPV, IPW, IPY, IQV, IQW, IQY, IRV, IRW, IRY, ISW, ISY, ITY, LCY, VFY, VWY
59	14	IDP, IEP, IGP, IHP, IKP, INP, IPP, IPQ, IPR, IPS, IPT, VPP, VPW, VPY
60	3	PKK, PKR, PRR
61	2	PRS, PRT
62	27	PDK, PDQ, PDR, PDS, PDT, PER, PGG, PGK, PGN, PGQ, PGR, PGS, PGT, PHK, PKN, PKQ, PKS, PKT, PNQ, PNR, PNS, PNT, PQQ, PQR, PQS, PSS, PST
63	5	HIK, HIR, HKV, HLR, HRV
64	3	HDG, HGN, NMM
65	6	HII, HIL, HIV, HLV, HVV, MRR
66	29	HDK, HDQ, HDR, HDS, HDT, HER, HGK, HGQ, HGR, HGT, HHK, HHR, HKN, HKQ, HKS, HKT, HNQ, HNR, HNS, HNT, HQQ, HQR, HQS, HQT, HRS, HRT, HSS, HST, QCH
67	120	HAC, HAF, HAH, HAI, HAK, HAM, HAN, HAP, HAQ, HAR, HAS, HAT, HAV, HAW, HAY, HCD, HCE, HCF, HCG, HCH, HCI, HCK, HCL, HCM, HCN, HCP, HCQ, HCR, HCS, HCT, HCV, HCW, HCY, HDD, HDF, HDH, HDI, HDL, HDM, HDN, HDP, HDV, HDW, HDY, HEI, HEP, HET, HEV, HEY, HFG, HFH, HFN, HFP, HFQ, HFR, HFS, HFT, HFV, HGH, HGM, HGW, HGY, HHH, HHI, HHL, HHM, HHN, HHP, HHQ, HHS, HHT, HHV, HHW, HHY, HIN, HIP, HIQ, HIT, HIY, HKP, HKY, HLN, HLP, HLQ, HLT, HLY, HMM, HMN, HMP, HMR, HMS, HMT, HMV, HMY, HNN, HNP, HNV, HNW, HNY, HPQ, HPR, HPT, HPV, HPW, HPY, HQY, HRY, HSW, HSY, HTT, HTW, HTY, HVY, HYY, QCC, QCF, QCM, QCW, QHW, QMM
68	13	HFF, HFI, HFL, HFM, HFW, HFY, HIM, HIW, HLW, HMW, HVW, HWW, HWY
69	8	HGI, HGL, HGV, HIS, HLS, HQV, HSV, HTV
70	14	HEF, HEL, HEM, HEW, HFK, HKL, HKM, HKW, HMQ, HQW, HRW, QFW, QMW, QWW
71	11	HAD, HAE, HAG, HDE, HEE, HEG, HEH, HEK, HEN, HEQ, HES
72	13	WAD, WAE, WAG, WDD, WDE, WDG, WEE, WEG, WEH, WEK, WEN, WEQ, WES
73	9	FWW, WFI, WFL, WFV, WII, WIL, WIV, WLL, WLV
74	6	WIK, WIR, WKL, WKV, WLR, WRV
75	10	WAP, WGI, WGL, WGV, WIS, WIT, WLS, WSV, WTT, WTV
76	13	WCC, WCD, WCF, WCG, WCH, WCI, WCM, WCN, WCQ, WCR, WCS, WCT, WCV
77	10	WDP, WEP, WGP, WKP, WNP, WPP, WPQ, WPR, WPS, WPT
78	108	WAC, WAF, WAH, WAI, WAK, WAL, WAM, WAN, WAQ, WAR, WAS, WAT, WAW, WAY, WCE, WCK, WCL, WCP, WCW, WCY, WDF, WDH, WDI, WDL, WDM, WDN, WDV, WDW, WDY, WEF, WEI, WEL, WEM, WET, WEV, WEW, WEY, WFF, WFH, WFK, WFM, WFN, WFQ, WFR, WFS, WFT, WFW, WFY, WGH, WGM, WGY, WHH, WHI, WHL, WHM, WHN, WHP, WHQ, WHR, WHS, WHT, WHV, WHW, WHY, WIM, WIN, WIQ, WIW, WIY, WKM, WKW, WKY, WLM, WLN, WLQ, WLT, WLW, WLY, WMM, WMN, WMP, WMQ, WMR, WMS, WMT, WMV, WMW, WMY, WNN, WNV, WNW, WNY, WPY, WQT, WQV, WQW, WQY, WRW, WRY, WSW, WSY, WTW, WTY, WVW, WVY, WWW, WWY, WYY
79	7	WFG, WFP, WGW, WIP, WLP, WPV, WPW
80	4	ACC, ACI, ACM, ACV
81	7	ACD, ACG, ACH, ACN, ACQ, ACS, ACT
82	7	CHR, CKR, CNR, CQR, CRR, CRS, CRT
83	25	CDK, CDQ, CDR, CDS, CDT, CER, CGK, CGQ, CGR, CGS, CGT, CHK, CKK, CKN, CKQ, CKS, CKT, CNQ, CNS, CNT, CQQ, CQS, CQT, CSS, CST
84	13	RFF, RFM, RFW, RFY, RIM, RIW, RIY, RLM, RLW, RMV, RVW, RWW, RWY
85	8	RFI, RFL, RFV, RII, RIL, RIV, RLL, RLV

continued on next page.

continued from previous page.

cluster #	# of the members	members in the cluster #
86	114	TWW, YAA, YAC, YAE, YAF, YAH, YAI, YAK, YAL, YAM, YAN, YAP, YAQ, YAR, YAS, YAT, YAV, YAW, YAY, YCE, YCF, YCI, YCK, YCL, YCM, YCP, YCV, YCW, YCY, YDF, YDI, YDL, YDM, YDV, YDW, YDY, YEF, YEI, YEL, YEM, YET, YEV, YEW, YEY, YFG, YFH, YFK, YFN, YFP, YFQ, YFR, YFS, YFT, YGH, YGI, YGL, YGM, YGV, YGW, YGY, YHH, YHI, YHL, YHM, YHN, YHP, YHQ, YHT, YHV, YHW, YHY, YIN, YIP, YIQ, YIS, YIY, YKL, YKM, YKW, YKY, YLN, YLP, YLQ, YLY, YMM, YMN, YMP, YMQ, YMR, YMS, YMT, YMV, YMY, YNV, YNW, YNY, YPT, YPV, YPW, YPY, YQV, YQW, YQY, YRW, YRY, YSV, YSW, YSY, YTT, YTW, YTY, YVY, YWY, YYY
87	9	YII, YIL, YIT, YIV, YLS, YLT, YLV, YTV, YVV
88	9	YDP, YEP, YGP, YKP, YNP, YPP, YPQ, YPR, YPS
89	9	YCC, YCD, YCG, YCH, YCN, YCQ, YCR, YCS, YCT
90	34	YDK, YDR, YDS, YDT, YER, YGK, YGN, YGQ, YGR, YGS, YGT, YHK, YHR, YHS, YKK, YKN, YKQ, YKR, YKS, YKT, YNN, YNQ, YNR, YNS, YNT, YQQ, YQR, YQS, YQT, YRR, YRS, YRT, YSS, YST
91	16	YAD, YAG, YDD, YDE, YDG, YDH, YDN, YDQ, YEE, YEG, YEH, YEK, YEN, YEQ, YES, YGG
92	15	YFF, YFI, YFL, YFM, YFV, YFW, YFY, YIM, YIW, YLL, YLM, YLW, YMW, YVW, YWW
93	5	YIK, YIR, YKV, YLR, YRV
94	5	NII, NIV, NVV, SHW, SMW
95	21	GRR, SDD, SDE, SDG, SDK, SDN, SDP, SDQ, SEG, SEK, SEN, SEP, SEQ, SER, SES, SGK, SGN, SKP, SNN, SNP, SPS
96	7	THR, TKR, TNR, TQR, TRR, TRS, TRT
97	20	SII, SIL, SIV, SLL, SLV, SVV, TAI, TAL, TAT, TAV, TGI, TGL, TGV, TIS, TIT, TLS, TSV, TTT, TTV, TVV
98	3	QGV, TEH, TEM
99	26	SDH, SDR, SDS, SDT, SGH, SGQ, SGR, SGS, SGT, SHK, SHQ, SHS, SKK, SKN, SKQ, SKS, SKT, SNQ, SNS, SNT, SQQ, SQS, SQT, SSS, SST, STT
100	139	NLV, NTV, SAA, SAC, SAD, SAE, SAF, SAG, SAH, SAI, SAK, SAL, SAM, SAN, SAP, SAQ, SAR, SAS, SAT, SAV, SAW, SAY, SCE, SCL, SCP, SCW, SCY, SDF, SDI, SDL, SDM, SDV, SDY, SEF, SEH, SEI, SEL, SEM, SET, SEV, SEY, SFG, SFH, SFK, SFN, SFP, SFQ, SFR, SFS, SFT, SGI, SGL, SGM, SGV, SGW, SGY, SHH, SHI, SHL, SHM, SHN, SHP, SHR, SHT, SHV, SHY, SIN, SIP, SKL, SKM, SKW, SKY, SLM, SLN, SLP, SLQ, SLS, SLT, SLY, SMM, SMN, SMP, SMQ, SMS, SMT, SMY, SNW, SNY, SPQ, SPR, SPT, SPV, SPW, SPY, SQW, SQY, SRW, SRY, SSW, SSY, STW, STY, SWY, SYY, TAD, TAE, TAG, TAK, TAN, TAP, TAS, TDD, TDE, TDG, TDH, TDN, TDP, TEE, TEG, TEK, TEN, TEP, TEQ, TES, TET, TGG, TGH, TGN, TGP, THN, THP, TKP, TNN, TNP, TPP, TPQ, TPR, TPS, TPT
101	26	SCC, SCD, SCF, SCG, SCH, SCI, SCK, SCM, SCN, SCQ, SCR, SCS, SCT, SCV, TCC, TCD, TCF, TCG, TCH, TCI, TCM, TCN, TCQ, TCS, TCT, TCV
102	6	SKR, SNR, SQR, SRR, SRS, SRT
103	101	SFF, SFI, SFL, SFM, SFV, SFW, SFY, SIM, SIW, SIY, SLW, SMV, SVW, SVY, TAA, TAC, TAF, TAH, TAM, TAQ, TAR, TAW, TAY, TCE, TCK, TCL, TCP, TCR, TCY, TDF, TDI, TDL, TDM, TDV, TDW, TDY, TEF, TEI, TEL, TEV, TEW, TEY, TFG, TFH, TFK, TFN, TFP, TFQ, TFR, TFS, TFT, TGM, TGW, TGY, THH, THI, THL, THM, THQ, THT, THV, THW, THY, TIK, TIN, TIP, TIQ, TIR, TKL, TKM, TKV, TKW, TKY, TLN, TLP, TLQ, TLR, TLT, TMN, TMP, TMQ, TMR, TMS, TMT, TNV, TNW, TNY, TPV, TPW, TPY, TQV, TQW, TQY, TRV, TRW, TRY, TSW, TSY, TTW, TTY, TYY
104	39	SIK, SIQ, SIR, SIS, SIT, SKV, SLR, SMR, SNV, SQV, SRV, SSV, STV, TDK, TDQ, TDR, TDS, TDT, TER, TGK, TGQ, TGR, TGS, TGT, THK, THS, TKK, TKN, TKQ, TKS, TKT, TNQ, TNS, TNT, TQQ, TQS, TQT, TSS, TST
105	23	TCW, TFF, TFI, TFL, TFM, TFV, TFW, TFY, TIM, TIW, TIY, TLL, TLM, TLV, TLW, TLY, TMM, TMV, TMW, TMY, TVW, TVY, TWY
106	12	PCC, PCD, PCF, PCG, PCH, PCI, PCM, PCN, PCQ, PCS, PCT, PCV
107	15	PAD, PAE, PAG, PAN, PDD, PDE, PDG, PDN, PEE, PEG, PEH, PEK, PEN, PEQ, PES
108	21	PAA, PAI, PAL, PAP, PAS, PAT, PAV, PGI, PGL, PGV, PGY, PIP, PIS, PIT, PLS, PPT, PPV, PPY, PSV, PTT, PTV
109	13	PFF, PFI, PFM, PFW, PFY, PIM, PIW, PLM, PLW, PMW, PVW, PWW, PWY
110	9	PDP, PEP, PGP, PKP, PNP, PPP, PPQ, PPR, PPS
111	97	PAC, PAF, PAH, PAK, PAM, PAQ, PAR, PAW, PAY, PCE, PCK, PCL, PCP, PCR, PCW, PCY, PDF, PDH, PDI, PDL, PDM, PDV, PDW, PDY, PEF, PEI, PEL, PEM, PET, PEV, PEW, PEY, PFG, PFH, PFK, PFN, PFQ, PFR, PFS, PFT, PGH, PGM, PGW, PHH, PHI, PHL, PHM, PHN, PHP, PHQ, PHR, PHS, PHT, PHV, PHW, PHY, PIK, PIN, PIQ, PIR, PIY, PKL, PKM, PKV, PKW, PKY, PLN, PLQ, PLR, PLT, PLY, PMM, PMN, PMP, PMQ, PMR, PMS, PMT, PMV, PMY, PNN, PNV, PNW, PNY, PQT, PQV, PQW, PQY, PRV, PRW, PRY, PSW, PSY, PTW, PTY, PVY, PYY
112	3	PFP, PLP, PPW
113	8	PFL, PFV, PII, PIL, PIV, PLL, PLV, PVV
114	6	AIK, AIR, AKL, AKV, ALR, ARV
115	5	AFV, AII, AIL, AIV, ALV
116	19	ACP, ADW, ADY, AGW, AHP, AHW, AHY, AKW, AKY, ANW, ANY, AQW, AQY, ARW, ARY, ASW, ASY, ATW, ATY
117	79	AAC, AAD, AAE, AAF, AAH, AAK, AAM, AAN, AAQ, AAR, AAS, AAW, AAY, ACE, ACK, ACL, ACR, ADD, ADE, ADF, ADH, ADI, ADL, ADM, ADN, ADV, AEE, AEF, AEG, AEH, AEI, AEK, AEL, AEM, AEN, AEP, AEQ, AES, AET, AEV, AEW, AEY, AFH, AFK, AFN, AFQ, AFR, AFS, AFT, AGH, AGM, AHH, AHI, AHL, AHM, AHN, AHQ, AHR, AHS, AHT, AHV, AIN, AIQ, AKM, ALN, ALP, ALQ, AMM, AMN, AMP, AMQ, AMR, AMS, AMT, AMV, AMY, ANN, ANV, AQV
118	12	AAI, AAL, AAT, AAV, AIS, AIT, ALS, ALT, ASV, ATT, ATV, AVV

continued on next page.

continued from previous page.

cluster #	# of the members	members in the cluster #
119	16	AFF, AFI, AFL, AFM, AFW, AFY, AIM, AIW, ALL, ALM, ALW, ALY, AMW, AVW, AWW, AWY
120	4	ADP, AKP, APQ, APR
121	6	ACF, ACW, ACY, AIY, AVY, AYY
122	34	RAA, RAY, RDI, RDL, REI, REL, REV, REW, RFG, RFK, RFN, RFP, RFS, RGI, RGL, RGM, RGV, RGW, RIP, RKL, RKM, RKW, RLN, RLP, RLQ, RLS, RMP, RNW, RPV, RPW, RPY, RQW, RRW, RSW
123	58	RAC, RAR, RAT, RCD, RCE, RCF, RCG, RCH, RCK, RCL, RCM, RCN, RCP, RCQ, RCR, RCW, RCY, RDV, RDY, RFQ, RFR, RFT, RGY, RHH, RHI, RHN, RHQ, RHT, RHV, RHY, RIK, RIN, RIQ, RIR, RIS, RIT, RKV, RKY, RLR, RLT, RLY, RMN, RMQ, RMR, RMS, RMT, RMY, RNV, RNY, RQV, RQY, RRV, RRY, RSY, RTW, RTY, RVY, RYY
124	3	RAI, RAL, RAV
125	9	RCC, RCI, RCS, RCT, RCV, RSV, RTT, RTV, RVV
126	23	RAE, RAF, RAH, RAM, RAW, RDD, RDE, RDF, RDH, RDM, RDN, RDW, REF, REH, REM, REN, REY, RFH, RHL, RHM, RHW, RMM, RMW
127	12	RAP, RDP, REP, RGP, RHP, RKP, RNP, RPP, RPQ, RPR, RPS, RPT
128	15	KCR, QKR, QRR, RAD, RAG, RAK, RAN, RAQ, RAS, REE, REG, REK, REQ, RES, RET
129	38	RDG, RDK, RDQ, RDR, RDS, RDT, RER, RGG, RGH, RGK, RGN, RGQ, RGR, RGS, RGT, RHK, RHR, RHS, RKK, RKN, RKQ, RKR, RKS, RKT, RNN, RNQ, RNR, RNS, RNT, RQQ, RQR, RQS, RQT, RRR, RRS, RRT, RSS, RST
130	7	LCI, LCL, LCV, LIT, LIV, LTV, LVV
131	12	IAE, IEE, LCD, LCG, LCH, LCK, LCN, LCQ, LCR, LCS, LCT, LHH
132	18	LFF, LFI, LFL, LFV, LFY, LII, LIL, LIM, LIW, LIY, LLL, LLM, LLV, LLW, LLY, LMV, LVW, LVY
133	11	LDW, LFH, LFM, LFW, LHW, LMW, LMY, LNW, LQW, LWW, LWY
134	3	LCF, LCM, LCW
135	41	LDG, LDK, LDN, LDQ, LDR, LDS, LDT, LER, LGG, LGH, LGK, LGN, LGQ, LGR, LGS, LGT, LHK, LHN, LHQ, LHR, LHS, LKK, LKN, LKQ, LKR, LKS, LKT, LNN, LNQ, LNR, LNS, LNT, LQQ, LQR, LQS, LQT, LRR, LRS, LRT, LSS, LST
136	82	LAA, LAC, LAF, LAG, LAI, LAL, LAP, LAR, LAS, LAT, LAV, LAW, LAY, LCP, LDF, LDI, LDL, LDV, LDY, LEI, LEL, LEV, LEW, LEY, LFG, LFK, LFN, LFP, LFQ, LFR, LFS, LFT, LGI, LGL, LGM, LGV, LGW, LGY, LHI, LHL, LHP, LHT, LHV, LHY, LIK, LIN, LIP, LIQ, LIR, LIS, LKL, LKM, LKV, LKW, LKY, LLN, LLP, LLQ, LLR, LLS, LLT, LMP, LMR, LMS, LMT, LNV, LNY, LPV, LPW, LPY, LQV, LQY, LRV, LRW, LRY, LSV, LSW, LSY, LTT, LTW, LTY, LYY
137	10	LDP, LEP, LGP, LKP, LNP, LPP, LPQ, LPR, LPS, LPT
138	26	LAD, LAE, LAH, LAK, LAM, LAN, LAQ, LCE, LDD, LDE, LDH, LDM, LEE, LEF, LEG, LEH, LEK, LEM, LEN, LEQ, LES, LET, LHM, LMM, LMN, LMQ
139	43	DRY, NDG, NDK, NDN, NDP, NDQ, NDR, NDS, NDT, NEG, NEK, NEN, NEP, NEQ, NER, NES, NGG, NGH, NGK, NGN, NGP, NGQ, NGR, NGS, NGT, NHK, NHR, NKN, NKQ, NKS, NNN, NNP, NNQ, NNR, NNS, NNT, NPQ, NQQ, NQR, NQS, NRS, NSS, NST
140	2	NCC, NCV
141	21	NIK, NIR, NKK, NKL, NKR, NKT, NKV, NKY, NLR, NRR, NRT, NRV, NRY, QDP, QGY, QNP, QPP, QPS, QPT, QPW, QPY
142	29	NAA, NAG, NAL, NAP, NAS, NAT, NAV, NFG, NFP, NGI, NGL, NGV, NGY, NIP, NIS, NLP, NLS, NPP, NPS, NPT, NPV, NPW, NPY, NSV, NSY, NTT, QGG, SDW, SEE
143	6	DYY, HPP, NFW, NHW, NMW, NWW
144	2	DRV, QDG
145	115	DTV, DVV, DVY, HGG, HGP, HGS, HPS, NAC, NAD, NAF, NAH, NAI, NAK, NAM, NAN, NAQ, NAR, NAW, NAY, NCE, NCF, NCI, NCL, NCM, NCP, NCW, NCY, NDD, NDF, NDH, NDI, NDL, NDM, NDV, NDW, NDY, NEI, NET, NEV, NEY, NFF, NFH, NFI, NFK, NFL, NFM, NFN, NFQ, NFR, NFS, NFT, NFV, NFY, NGM, NGW, NHI, NHL, NHM, NHN, NHP, NHQ, NHS, NHT, NHV, NHY, NIL, NIM, NIN, NIQ, NIT, NIW, NIY, NKM, NKP, NKW, NLL, NLM, NLN, NLQ, NLT, NLW, NLY, NMN, NMP, NMQ, NMR, NMS, NMT, NMV, NMY, NNV, NNW, NNY, NPR, NQT, NQV, NQW, NQY, NRW, NSW, NTW, NTY, NVW, NVY, NWY, NYY, QCG, QCP, QGN, QGP, QGS, QGT, QGW, SEW, SWW
146	10	NCD, NCG, NCH, NCK, NCN, NCQ, NCR, NCS, NCT, NHH
147	17	DAA, DDF, DDW, DEF, DEL, DEM, DEW, DFG, DFP, DGL, DGW, DKW, DLP, DMP, DNW, DPW, DPY
148	141	DAC, DAD, DAF, DAG, DAH, DAK, DAM, DAN, DAP, DAQ, DAR, DAS, DAT, DAW, DAY, DCE, DCG, DCK, DCL, DCP, DCR, DCY, DDD, DDG, DDH, DDI, DDK, DDL, DDM, DDN, DDQ, DDR, DDS, DDT, DDV, DDY, DEG, DEH, DEI, DEK, DEN, DEQ, DER, DES, DET, DEV, DEY, DFH, DFK, DFN, DFQ, DFR, DFS, DFT, DGH, DGI, DGK, DGM, DGN, DGQ, DGR, DGT, DGV, DGY, DHH, DHI, DHK, DHL, DHM, DHN, DHP, DHQ, DHR, DHS, DHT, DHV, DHW, DHY, DIK, DIN, DIP, DIQ, DIR, DIS, DKK, DKL, DKM, DKN, DKQ, DKR, DKS, DKT, DKV, DKY, DLN, DLQ, DLR, DLS, DMN, DMQ, DMR, DMS, DMT, DMV, DMY, DNN, DNQ, DNR, DNS, DNT, DNV, DNY, DPQ, DPR, DPT, DPV, DQQ, DQR, DQS, DQT, DQV, DQW, DQY, DRS, DRT, DRW, DSS, DST, DSV, DSW, DSY, DTT, DTW, DTY, NAE, NDE, NEE, NEF, NEH, NEM, NEW
149	3	DAE, DDE, DEE
150	11	DCC, DCD, DCF, DCH, DCI, DCM, DCN, DCQ, DCS, DCT, DCV
151	9	DDP, DEP, DGG, DGP, DGS, DKP, DNP, DPP, DPS
152	13	DAI, DAL, DAV, DFI, DFV, DII, DIL, DIT, DIV, DLL, DLT, DLV, NEL
153	17	DCW, DFF, DFL, DFM, DFW, DFY, DIM, DIW, DIY, DLM, DLW, DLY, DMM, DMW, DVW, DWW, DWY
154	5	VII, VIL, VIV, VLV, VVV
155	43	HLL, MDD, MDG, MDH, MDK, MDN, MDQ, MDR, MDS, MDT, MEP, MER, MGG, MGH, MGK, MGN, MGQ, MGR, MGS, MGT, MHK, MHN, MHR, MHS, MKK, MKN, MKQ, MKR, MKS, MKT, MNN, MNQ, MNR, MNS, MNT, MQQ, MQR, MQS, MQT, MRS, MRT, MSS, MST

continued on next page.

continued from previous page.

cluster #	# of the members	members in the cluster #
156	13	MAG, MAH, MAK, MAN, MAQ, MAS, MEG, MEH, MEK, MEN, MEQ, MES, MET
157	13	MFI, MFV, MFY, MIL, MIM, MIW, MIY, MLL, MLY, MMV, MVW, MVY, MYY
158	8	MCC, MCF, MCI, MCL, MCM, MCV, MCW, MCY
159	28	HLM, MCP, MDP, MDY, MFP, MGP, MGW, MGY, MHP, MHY, MIP, MKP, MKW, MKY, MNP, MNY, MPQ, MPR, MPS, MPT, MQW, MQY, MRW, MRY, MSW, MSY, MTW, MTY
160	10	MAI, MAV, MII, MIT, MIV, MLS, MLT, MLV, MTV, MVV
161	65	MAC, MAP, MAR, MAT, MAY, MCD, MCE, MCG, MCH, MCK, MCN, MCQ, MCR, MCS, MCT, MDI, MDM, MDV, MEI, MEV, MEY, MFG, MFH, MFK, MFN, MFQ, MFR, MFS, MFT, MGI, MGL, MGM, MGV, MHH, MHI, MHL, MHM, MHQ, MHT, MHV, MIK, MIN, MIQ, MIR, MIS, MKL, MKM, MKV, MLN, MLP, MLQ, MLR, MMM, MMN, MMQ, MMR, MMS, MMT, MMY, MNV, MPV, MQV, MRV, MSV, MTT
162	2	HAA, HAL
163	8	AKT, ANR, ANT, AQR, AQT, ARS, ART, AST
164	3	AKK, AKR, ARR
165	23	ADG, ADK, ADQ, ADR, ADS, ADT, AER, AGG, AGK, AGN, AGQ, AGR, AGS, AGT, AHK, AKN, AKQ, AKS, ANQ, ANS, AQQ, AQS, ASS
166	3	AGI, AGL, AGV
167	3	AAA, AAG, AFG
168	12	AAP, AFP, AGP, AGY, AIP, ANP, APP, APS, APT, APV, APW, APY
169	7	QFG, QFP, QGI, QGL, QIP, QLP, QPV
170	35	KCC, KCG, KCH, QDH, QDQ, QDR, QDS, QDT, QGH, QGK, QGQ, QGR, QHK, QHN, QHQ, QHR, QHS, QKN, QKQ, QKS, QKT, QNN, QNQ, QNR, QNS, QNT, QQQ, QQR, QQS, QQT, QRS, QRT, QSS, QST, QTT
171	25	HKK, HKR, HRR, KCM, QDW, QEW, QFK, QFR, QHH, QHY, QIK, QIR, QKL, QKM, QKV, QKW, QKY, QLR, QMQ, QMR, QQW, QQY, QRV, QRW, QRY
172	2	QKP, QPR
173	37	DRR, QAA, QAD, QAE, QAF, QAG, QAH, QAK, QAM, QAN, QAP, QAQ, QAS, QDD, QDE, QDK, QDL, QDM, QDN, QEE, QEF, QEG, QEH, QEI, QEK, QEL, QEM, QEN, QEP, QEQ, QER, QES, QET, QEY, QGM, QHP, QPQ
174	50	QAC, QAR, QAW, QAY, QCD, QCE, QCI, QCK, QCL, QCN, QCQ, QCR, QCS, QCT, QCV, QCY, QDF, QDI, QDV, QDY, QEV, QFH, QFN, QFQ, QFS, QFT, QHI, QHL, QHM, QHT, QHV, QIN, QIQ, QLN, QLQ, QMN, QMP, QMS, QMT, QMY, QNV, QNW, QNY, QQV, QSW, QSY, QTW, QTY, QVY, QYY
175	14	QAI, QAL, QAT, QAV, QII, QIS, QIT, QIV, QLS, QLT, QLV, QSV, QTV, QVV
176	17	QFF, QFI, QFL, QFM, QFV, QFY, QIL, QIM, QIW, QIY, QLL, QLM, QLW, QLY, QMV, QVW, QWY
177	5	KII, KIL, KIV, KLV, KVV
178	84	KAC, KAF, KAH, KAK, KAM, KAQ, KAR, KAW, KAY, KCD, KCE, KCF, KCK, KCL, KCN, KCP, KCQ, KCW, KCY, KDF, KDH, KDI, KDL, KDM, KDV, KDW, KDY, KEF, KEI, KEL, KEM, KET, KEV, KEW, KEY, KFH, KFK, KFN, KFQ, KFR, KFS, KFT, KGM, KHH, KHI, KHL, KHM, KHN, KHP, KHQ, KHS, KHT, KHV, KHW, KHY, KIN, KIQ, KKM, KLN, KLQ, KLT, KLY, KMM, KMN, KMP, KMQ, KMR, KMS, KMT, KMV, KMY, KNV, KNW, KNY, KQV, KQW, KQY, KSW, KSY, KTW, KTY, KVY, KYY, QKK
179	14	KAD, KAE, KAG, KAN, KDD, KDE, KDG, KEE, KEG, KEH, KEK, KEN, KEQ, KES
180	3	KAI, KAL, KAV
181	37	KDK, KDN, KDQ, KDR, KDS, KDT, KER, KGG, KGH, KGK, KGN, KGQ, KGR, KGS, KGT, KHK, KHR, KKK, KKN, KKQ, KKR, KKS, KKT, KNN, KNQ, KNR, KNS, KNT, KQQ, KQR, KQS, KQT, KRR, KRS, KRT, KSS, KST
182	9	KCI, KCS, KCT, KCV, KIS, KIT, KSV, KTT, KTV
183	9	KDP, KEP, KGP, KKP, KNP, KPP, KPQ, KPR, KPS
184	4	KKW, KKY, KRW, KRY
185	17	KFF, KFI, KFL, KFM, KFV, KFW, KFY, KIM, KIW, KIY, KLL, KLM, KLW, KMW, KVW, KWW, KWY
186	18	KAA, KAP, KAS, KAT, KFG, KFP, KGI, KGL, KGV, KGW, KGY, KIP, KLP, KLS, KPT, KPV, KPW, KPY
187	6	KIK, KIR, KKL, KKV, KLR, KRV
188	10	FDP, FEP, FGP, FKP, FNP, FPP, FPQ, FPR, FPS, FPT
189	5	FFR, FKW, FKY, FRW, FRY
190	7	FDW, FEW, FEY, FHW, FQW, WAA, WAV
191	23	FAD, FAE, FAG, FAK, FAN, FAP, FAQ, FAS, FDD, FDE, FDG, FDK, FEE, FEG, FEH, FEK, FEN, FEQ, FER, FES, FET, FGK, FNN
192	11	FAI, FAL, FAV, FFV, FII, FIL, FIT, FIV, FLL, FLT, FLV
193	16	FAA, FFG, FFP, FFS, FGI, FGL, FGV, FGW, FIP, FLP, FLS, FMP, FPV, FPW, FPY, FSV
194	11	FCP, FDY, FGY, FHP, FHY, FNW, FNY, FQY, FSW, FSY, FTY
195	34	FAC, FAH, FAR, FAT, FAY, FCE, FDF, FDI, FDL, FDM, FDV, FEF, FEI, FEL, FEM, FEV, FFK, FFN, FFQ, FGM, FHM, FIK, FIN, FIQ, FKL, FKM, FKV, FLN, FLQ, FLR, FMQ, FMS, FMT, FQV
196	34	FAF, FAM, FAW, FCF, FCL, FCW, FCY, FFF, FFH, FFI, FFL, FFM, FFT, FFW, FFY, FHI, FHL, FHV, FIM, FIW, FIY, FLM, FLW, FLY, FMM, FMV, FMW, FMY, FTW, FVW, FVY, FWY, FYY, WVV
197	10	ECC, ECD, ECG, ECH, ECN, ECQ, ECR, ECS, ECT, EHH
198	36	ECK, EDH, EDQ, EDR, EDS, EDT, EGG, EGH, EGN, EGQ, EGR, EGS, EGT, EHK, EHN, EHQ, EHR, EHS, EHT, EKN, EKQ, EKS, EKT, EMR, ENQ, ENS, ENT, EQQ, EQR, EQS, EQT, ERS, ERT, ESS, EST, ETT
199	3	EKK, EKR, ERR
total		4200

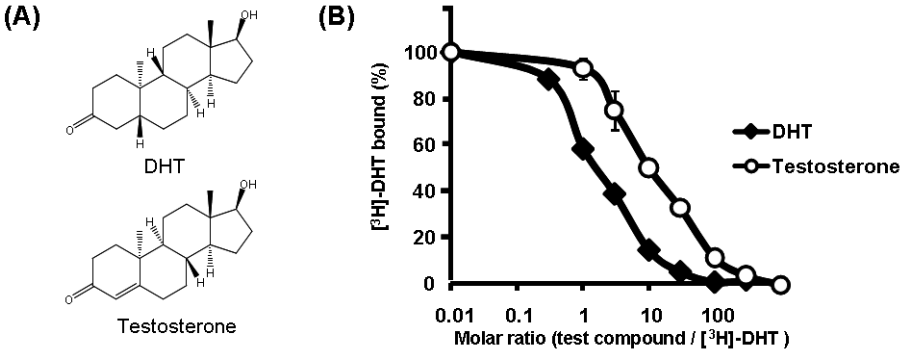
Supplementary Table A41 Predicted target proteins of MDMA. One general binding prediction model using C, F and G as feature vectors with 2,000 negatives was used.

	1 ID	description	2 probability
1	P08588	Beta-1 adrenergic receptor	0.956
2	P23975	Sodium-dependent noradrenaline transporter (NET)	0.930
3	P31645	Sodium-dependent serotonin transporter (5HTT)	0.930
4	P03372	Estrogen receptor (ER) (ER-alpha)	0.905
5	P35348	Alpha-1A adrenergic receptor	0.905
6	P35372	Mu-type opioid receptor (MOR-1)	0.902
7	P41145	Kappa-type opioid receptor (KOR-1)	0.899
8	P28223	5-hydroxytryptamine 2A receptor (5-HT-2A) (5-HT-2)	0.894
9	Q01959	Sodium-dependent dopamine transporter (DAT)	0.894
10	P35367	Histamine H1 receptor	0.885
11	P14416	D(2) dopamine receptor	0.879
12	P21728	D(1A) dopamine receptor	0.875
13	Q15822	Neuronal acetylcholine receptor protein, alpha-2 subunit precursor	0.873
14	P41143	Delta-type opioid receptor (DOR-1)	0.868
15	P11229	Muscarinic acetylcholine receptor M1	0.867
16	P08913	Alpha-2A adrenergic receptor (Alpha-2AAR)	0.847
17	P14867	Gamma-aminobutyric-acid receptor alpha-1 subunit precursor (GABA(A) receptor)	0.836
18	P41595	5-hydroxytryptamine 2B receptor (5-HT-2B)	0.832
19	Q06432	Voltage-dependent calcium channel gamma-1 subunit	0.830
20	Q8TCU5	Glutamate [NMDA] receptor subunit 3A precursor (NMDAR-L)	0.810
21	P28335	5-hydroxytryptamine 2C receptor (5-HT-2C) (5-HT2C) (5-HTR2C) (5HT-1C)	0.790
22	P28221	5-hydroxytryptamine 1D receptor (5-HT-1D) (5-HT-1D-alpha)	0.779
23	Q14524	Sodium channel protein type V alpha subunit (HH1)	0.747
24	P20309	Muscarinic acetylcholine receptor M3	0.736
25	P21917	D(4) dopamine receptor (D(2C) dopamine receptor)	0.711
26	P25100	Alpha-1D adrenergic receptor	0.698
27	*Q9UKG4	Solute carrier family 13 member 4 (Na(+)/sulfate cotransporter SUT-1)	0.678
28	P35368	Alpha-1B adrenergic receptor	0.669
29	P06858	Lipoprotein lipase precursor (LPL)	0.664
30	P08912	Muscarinic acetylcholine receptor M5	0.645
31	P22748	Carbonic anhydrase 4 precursor (CA-IV)	0.643
32	*Q8TCT7	Signal peptide peptidase-like 2B (IMP4)	0.611
33	P00915	Carbonic anhydrase 1 (CA-I)	0.588
34	P35354	Prostaglandin G/H synthase 2 precursor (Cyclooxygenase-2) (COX-2) (PGHS-2)	0.585
35	*Q99705	Melanin-concentrating hormone receptor 1 (MCHR-1) (MCH-R1) (MCH1R) (MCH-1R) (MCHR) (SLC-1)	0.583
36	O60779	Thiamine transporter 1 (THTR-1) (ThTr1) (TC1)	0.579
37	P21397	Amine oxidase [flavin-containing] A (MAO-A)	0.578
38	*Q92911	Sodium/iodide cotransporter (Na(+)/I(-) cotransporter) (Na(+)/I(-)-symporter)	0.570
39	*Q8N6U8	Probable G-protein coupled receptor 161	0.564
40	*O43614	Orexin receptor type 2 (Ox2r)	0.554
41	Q09470	Potassium voltage-gated channel subfamily A member 1 (HUKI) (HBK1)	0.552
42	*Q695T7	Sodium-dependent neutral amino acid transporter B(0) (B(0)AT1)	0.550
43	*P21730	C5a anaphylatoxin chemotactic receptor (C5a-R) (C5aR)	0.549
44	*Q8TE23	Taste receptor type 1 member 2 precursor	0.547
45	P98194	Calcium-transporting ATPase type 2C member 1	0.542
46	*P30872	Somatostatin receptor type 1 (SS1R) (SRIF-2)	0.540
47	*Q9NSD5	Sodium- and chloride-dependent GABA transporter 2	0.528
48	*P34969	5-hydroxytryptamine 7 receptor (5-HT-7) (5-HT-X) (5HT7)	0.524
49	Q9GZZ6	Neuronal acetylcholine receptor protein, alpha-10 subunit precursor	0.522
50	*Q9ULW2	Frizzled 10 precursor (Frizzled-10) (Fz-10) (hFz10) (FzE7)	0.521
51	*Q14832	Metabotropic glutamate receptor 3 precursor (mGluR3)	0.516
52	Q9Y5Y9	Sodium channel protein type X alpha subunit (hPN3)	0.515
53	Q8TDS4	Nicotinic acid receptor 1	0.500
54	*Q9UMX9	Membrane-associated transporter protein	0.500
55	*P31391	Somatostatin receptor type 4 (SS4R)	0.500
56	*Q9UN76	Sodium- and chloride-dependent neutral and basic amino acid transporter B(0+)	0.500

1: UniProt ID. 2: the estimated probability that the chemical compound binds to the protein. *: proteins that are not in the training data.

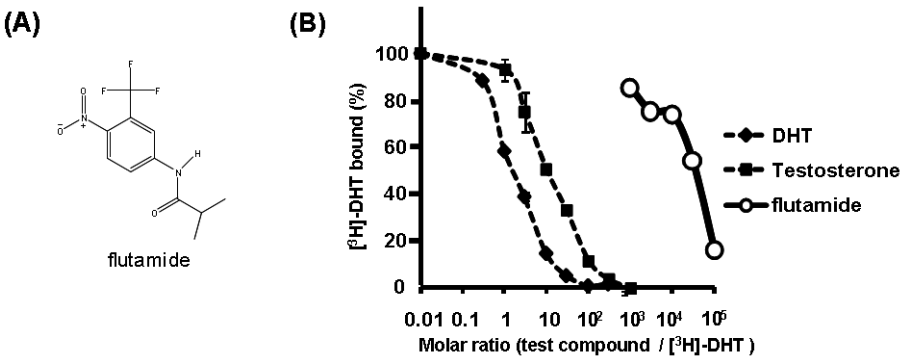
(First experimental verification)

I. *in vitro* binding assay of DHT and testosterone



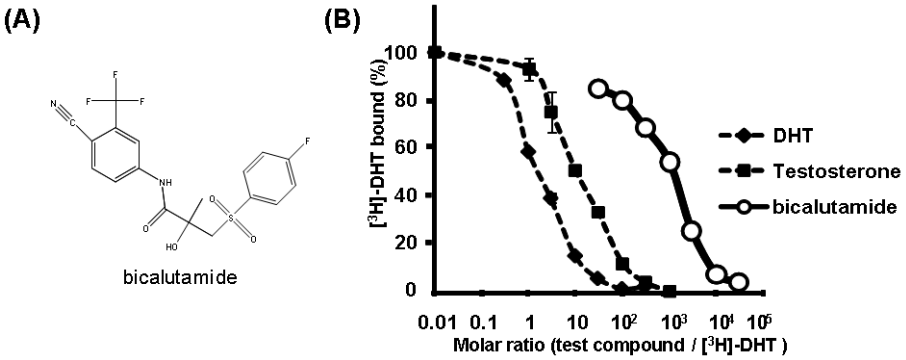
(A) chemical structure of dihydrotestosterone (DHT) and testosterone. (B) result of *in vitro* binding assay.

II. *in vitro* binding assay of flutamide



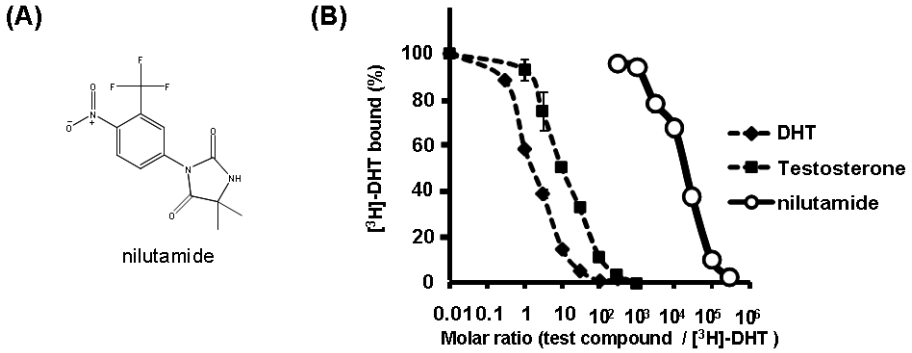
(A) chemical structure of flutamide. (B) result of *in vitro* binding assay.

III. *in vitro* binding assay of bicalutamide



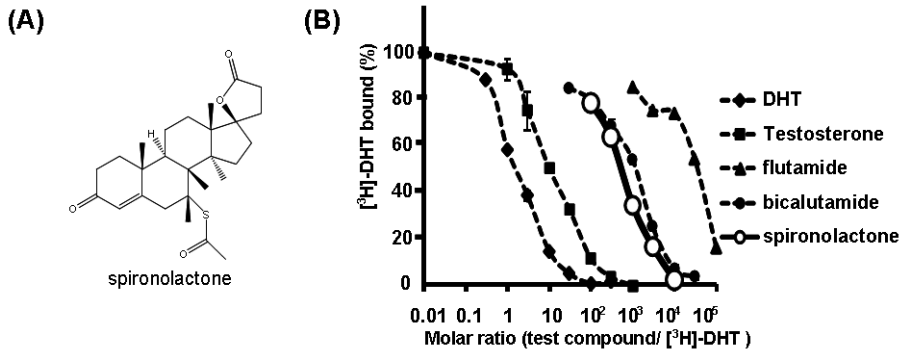
(A) chemical structure of bicalutamide. (B) result of *in vitro* binding assay.

IV. *in vitro* binding assay of nilutamide



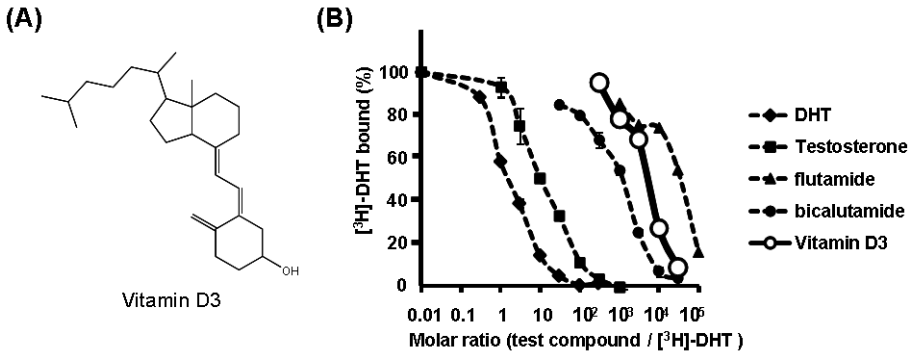
(A) chemical structure of nilutamide. (B) result of *in vitro* binding assay.

V. *in vitro* binding assay of spironolactone



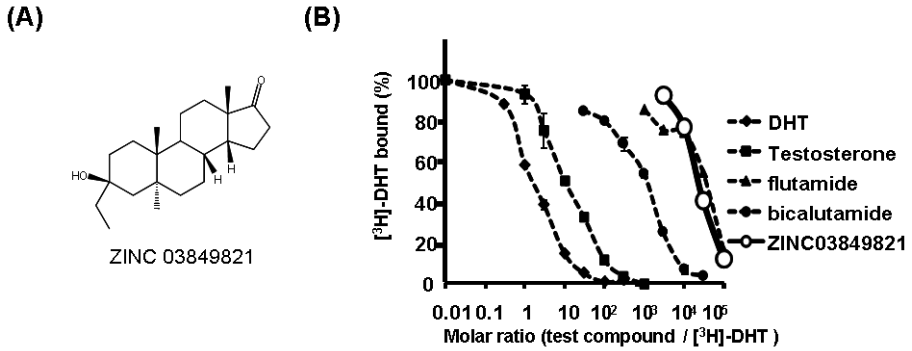
(A) chemical structure of spironolactone. (B) result of *in vitro* binding assay.

VI. *in vitro* binding assay of vitamin D3



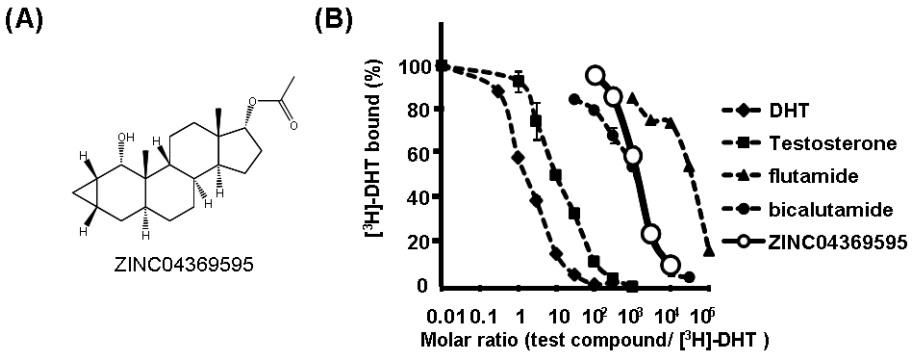
(A) chemical structure of vitamin D3. (B) result of *in vitro* binding assay.

VII. *in vitro* binding assay of ZINC 03849821



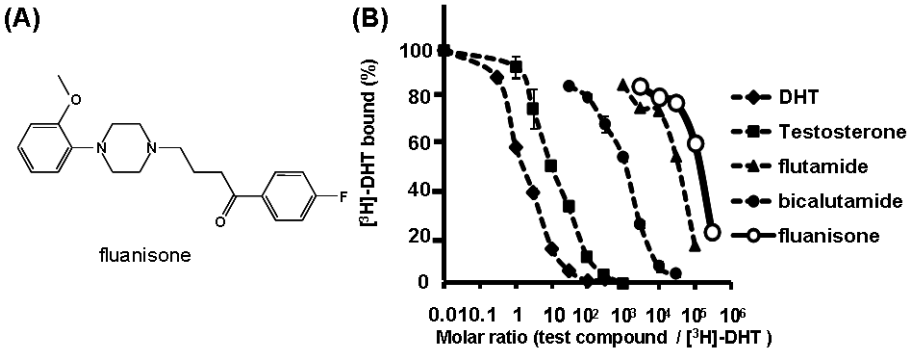
(A) chemical structure of ZINC 03849821. (B) result of *in vitro* binding assay.

VIII. *in vitro* binding assay of ZINC 04369595



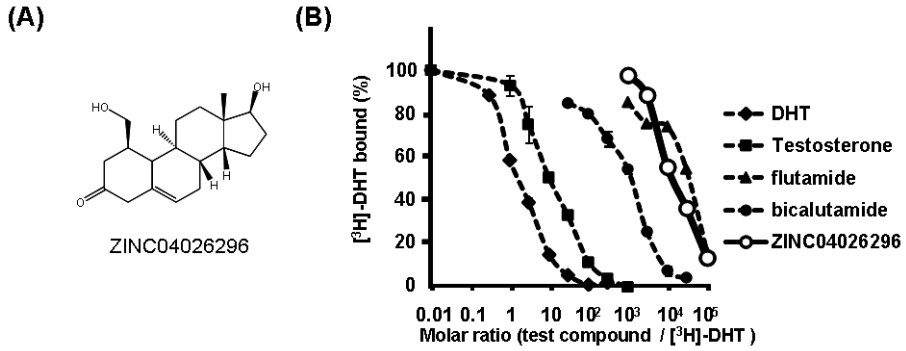
(A) chemical structure of ZINC 04369595. (B) result of *in vitro* binding assay.

IX. *in vitro* binding assay of fluanisone



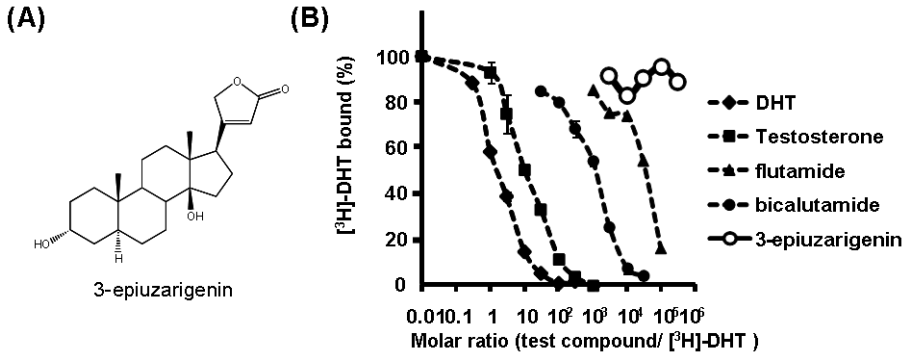
(A) chemical structure of fluanisone. (B) result of *in vitro* binding assay.

X. *in vitro* binding assay of ZINC 04026296



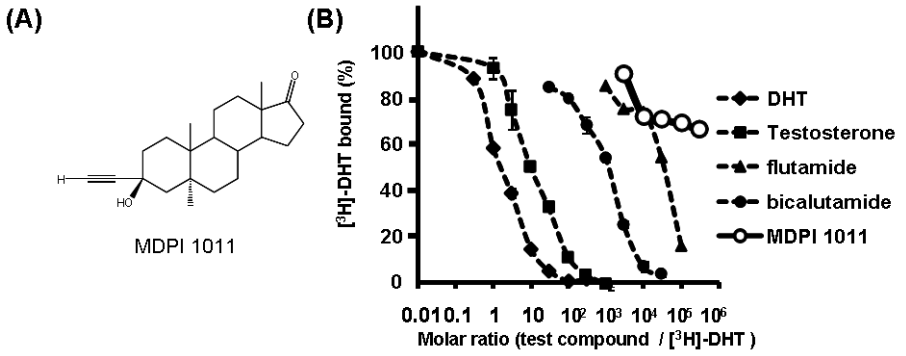
(A) chemical structure of ZINC 04026296. (B) result of *in vitro* binding assay.

XI. *in vitro* binding assay of 3-epiuzarigenin



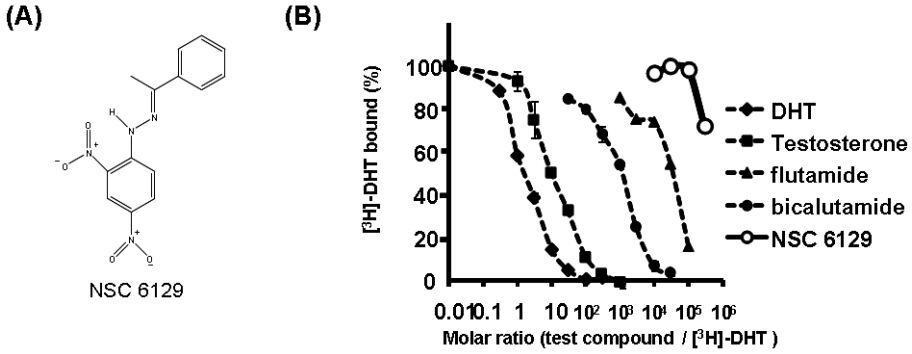
(A) chemical structure of 3-epiuzarigenin. (B) result of *in vitro* binding assay.

XII. *in vitro* binding assay of MDPI 1011



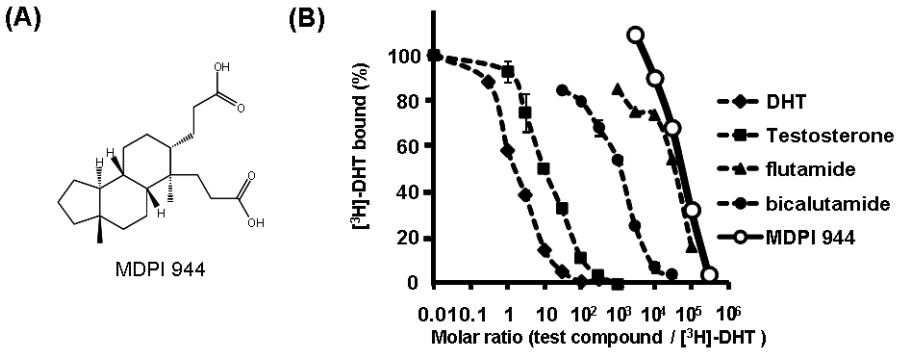
(A) chemical structure of MDPI 1011. (B) result of *in vitro* binding assay.

XIII. *in vitro* binding assay of NSC 6129



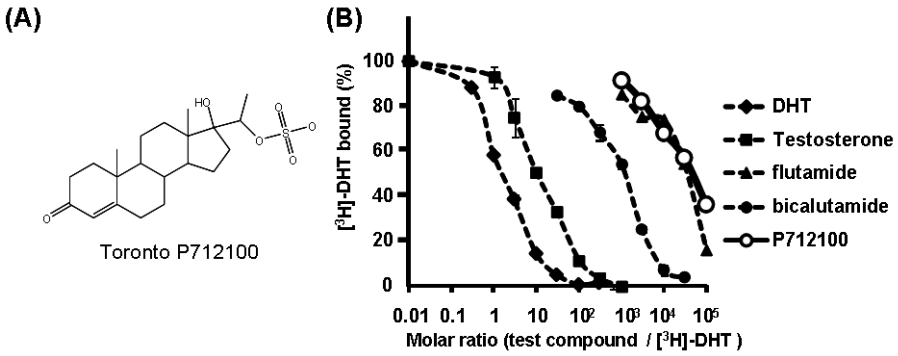
(A) chemical structure of NSC 6129. (B) result of *in vitro* binding assay.

XIV. *in vitro* binding assay of MDPI 944



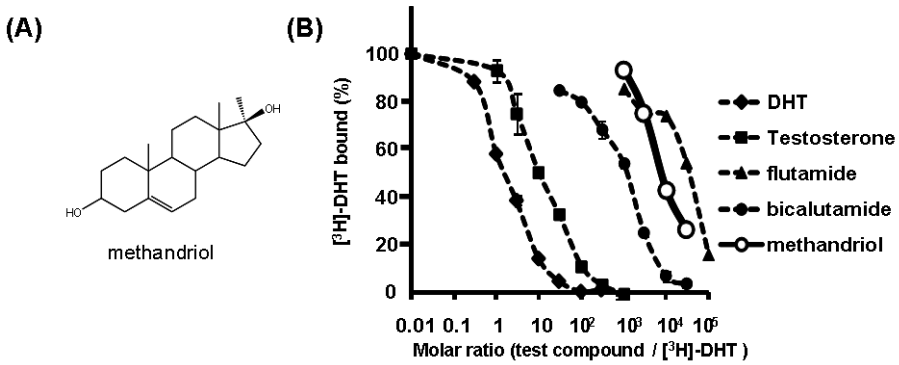
(A) chemical structure of MDPI 944. (B) result of *in vitro* binding assay.

XV. *in vitro* binding assay of Toronto P712100



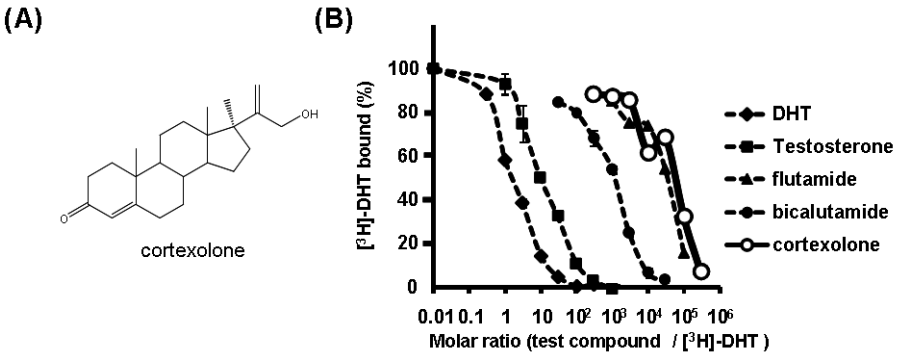
(A) chemical structure of Toronto P712100. (B) result of *in vitro* binding assay.

XVI. *in vitro* binding assay of methandriol



(A) chemical structure of methandriol. (B) result of *in vitro* binding assay.

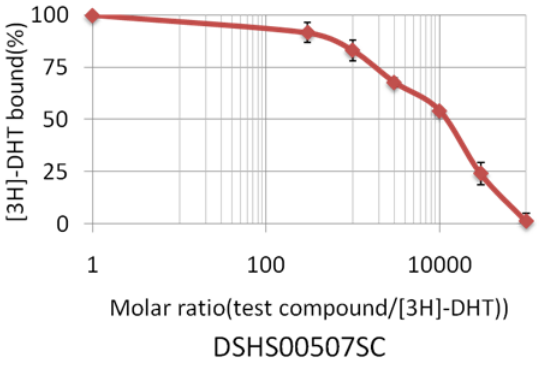
XVII. *in vitro* binding assay of cortexolone

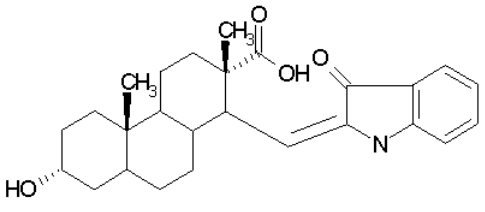


(A) chemical structure of cortexolone. (B) result of *in vitro* binding assay.

(Second experimental verification)

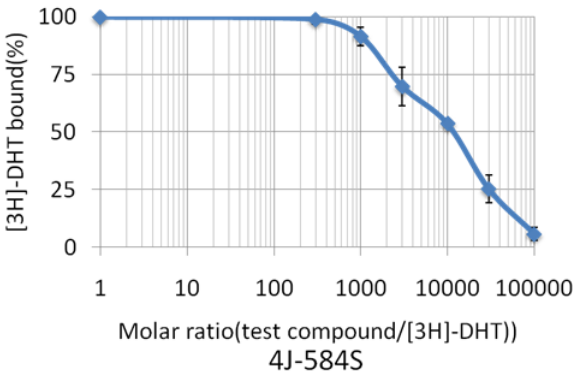
XVIII. *in vitro* binding assay of DSHS00507SC

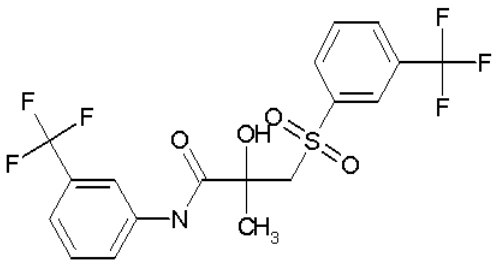




IC_{50} 22 μM

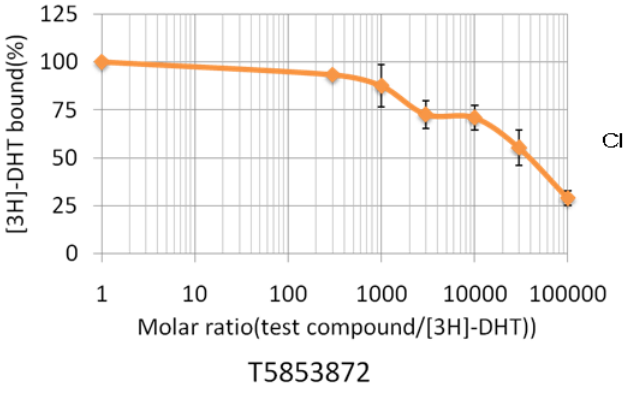
XIX. *in vitro* binding assay of 4J-584S

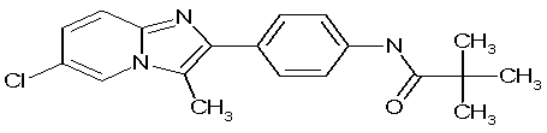




IC_{50} 22 μM

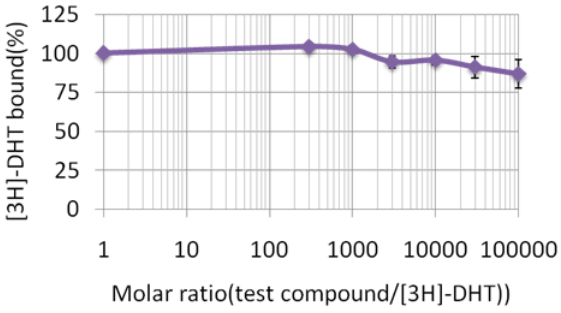
XX. *in vitro* binding assay of T5853872

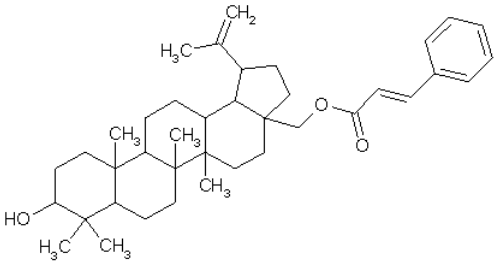




IC_{50} 80 μM

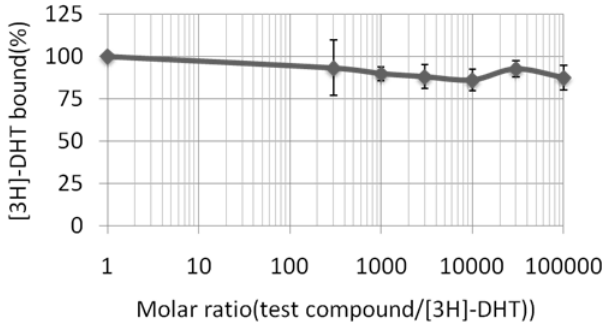
XXI. *in vitro* binding assay of BAS01279920

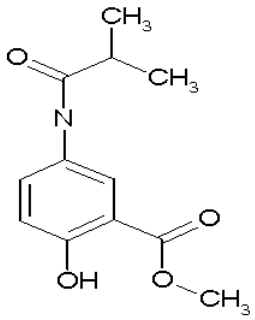




IC_{50}> 200μM

XXII. *in vitro* binding assay of AN-652/43163258





IC_{50}> 200μM

Supplementary Fig. 1 Result of *in vitro* binding assay for each compound

Supplementary Table A42 Details of predictions in Fig. III.16A

	[1] ID	[2] prob.	[3] H don.	[4] H accept.	[5] M.W.	xlogp	[6] Lipinski5	compounds of same connectivity
1	10839762	0.991	1	4	522.76	9.9	no	-
2	11364869	0.990	1	4	550.81	10.2	no	-
3	21188348	0.990	3	7	704.97	[7] N.A.	no	-
4	13132668	0.989	1	4	460.69	6.4	no	-
5	11314891	0.988	1	4	376.53	5	no	-
6	130408	0.988	2	4	346.46	1.8	yes	21121280
7	261801	0.987	1	4	430.62	6.3	no	-
8	23322680	0.987	1	4	446.66	6.9	no	-
9	11687	0.987	2	4	346.46	2.6	yes	627480 22296005
10	23322669	0.986	1	4	446.66	6.7	no	-
11	625566	0.986	3	4	348.48	2.6	yes	107483 160794 9902917 10427858
12	13132666	0.986	1	4	432.64	6.6	no	-
13	10504917	0.985	2	2	480.76	9.3	no	-
14	10382251	0.985	1	4	334.45	4	yes	-
15	21777241	0.985	1	5	502.73	6.7	no	-
16	5878	0.984	1	3	306.44	4.2	yes	249098 6429885 6432530 9883009 23616084
17	23724680	0.984	2	3	430.66	5.5	no	-
18	20317298	0.984	1	4	372.5	3.9	yes	-
19	22900385	0.983	4	10	979.41	14.6	no	-
20	155549	0.983	0	3	402.61	7.5	no	3084167
21	23322724	0.983	1	4	414.58	5.4	no	-
22	20713669	0.983	1	2	274.4	3.7	yes	-
23	20848952	0.982	4	7	402.5	N.A.	-	-
24	21393623	0.982	1	2	288.42	3.5	yes	22800529
25	13132678	0.982	1	4	432.64	6.4	no	-
26	11157467	0.982	2	4	578.86	N.A.	no	-
27	623740	0.982	2	3	304.42	1.8	yes	150968 1149869 1778898 1778899 1778901 6546032 11872739 11872740 16395893 22849351
28	125155	0.982	2	3	416.64	5.4	no	21120484
29	10096703	0.981	2	2	480.76	9.4	no	11754921
30	3084142	0.981	0	3	416.64	6.8	no	13014291 13014314
31	20697581	0.980	2	5	538.76	7	no	-
32	19609992	0.980	2	3	304.42	2.6	yes	-
33	21850917	0.980	1	3	430.66	6.9	no	-
34	248854	0.979	2	3	412.6	6.7	no	-
35	65545	0.979	2	3	304.42	3	yes	237641 9904576 10086383 15708880 21139669 23624176
36	14259153	0.979	2	3	304.42	1.9	yes	23256526
37	22812420	0.978	1	4	522.76	8.4	no	23338816
38	17896827	0.977	2	2	402.65	7.3	no	-
39	21709398	0.977	2	3	318.45	2.6	yes	-
40	5284278	0.977	1	2	400.64	6.3	no	-
41	3733299	0.976	2	4	329.43	2.5	yes	16396395
42	536509	0.976	1	4	376.53	4	yes	-
43	10204	0.976	1	2	288.42	3.6	yes	5408 6013 186663 242356 369323 638016 711503 908606 908607 908608 1312122 2825665 3034651 5701998 6432567 6541850 7048589 10063008 10108190 10469710 11514789 11861311 11862155 11869421 11869422 11869423 11870497 11890585 11890586 11920270 12304539 12304540 12304570 16220011 16758160 19615208 20848917 22876164
44	21532454	0.976	1	3	299.24	N.A.	-	-
45	20317338	0.976	1	4	372.5	4.3	yes	-
46	2748175	0.975	1	2	288.42	3.1	yes	3639419 11886958 11886959 11886960 11886961
47	11433605	0.975	1	4	629.71	10.6	no	-
48	190885	0.975	1	2	400.64	6.2	no	-
49	11038956	0.974	1	4	553.61	7.4	no	21673496
50	22082233	0.973	2	6	504.7	7.3	no	-
51	20372517	0.971	1	3	372.54	4.4	yes	22801536
52	352575	0.969	2	3	290.4	1.7	yes	-
53	526978	0.968	0	3	402.61	6.5	no	160230 6428249 6428257 6428266 6428275 13267935 20837721
54	11797004	0.968	2	4	432.64	5.6	no	21585485 21774477
55	10648350	0.967	1	2	478.75	9	no	-
56	6428271	0.967	0	3	360.53	6.1	no	6428245 6428254 6428262 20837712 20837718
57	23322711	0.967	1	4	418.61	6	no	-
58	13132674	0.966	1	4	404.58	5.6	no	-
59	17896838	0.965	2	2	402.65	7.5	no	-
60	19027438	0.965	3	5	362.46	1.4	yes	-
61	10001994	0.964	2	2	416.68	8	no	-

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
62	16937	0.963	0	3	414.62	6.7	no	18186230
63	20743369	0.963	2	4	441.65	4.9	yes	23504634
64	630661	0.963	2	4	336.44	2	yes	-
65	150854	0.963	1	2	288.42	3.8	yes	10039774 10085744 14324931 21122962
66	251593	0.962	2	3	304.42	2.9	yes	4038155
67	17896824	0.962	2	2	402.65	7.3	no	-
68	248649	0.960	2	3	304.42	3.5	yes	3994025
69	14197	0.960	1	4	372.5	3.3	yes	1715111 1715112 1758149 1758150 3823590 11875333 11875334 11875335 11875336 16399658 23963785
70	4535820	0.958	1	5	572.84	8.7	no	-
71	11420412	0.958	2	5	488.7	6.4	no	21580207
72	10131684	0.958	2	2	332.52	5	no	-
73	10896648	0.958	1	4	522.76	8.2	no	-
74	3397	0.957	1	6	276.21	2.6	yes	10802904
75	10297173	0.957	0	3	442.67	7.3	no	-
76	20531740	0.957	2	3	332.48	3.8	yes	22820966
77	2375	0.957	2	9	430.37	2.5	yes	56069 441405 16110572
78	3170	0.957	0	3	360.53	5.8	no	224004 10316496 16627147 20056767 23615700
79	9986020	0.957	1	3	599.75	N.A.	no	-
80	3387	0.957	2	4	336.44	2	yes	6446 6429867 6432701 10088392 20056778
81	4435	0.957	0	3	406.56	5.6	no	229455 11869560 11872989 16394511 16396215 23615856
82	13132652	0.956	1	4	418.61	6.1	no	-
83	23322662	0.956	1	4	400.55	4.9	yes	-
84	18548013	0.955	1	2	288.42	N.A.	-	-
85	23495	0.955	2	3	352.9	2.1	yes	235091 3427265
86	19603132	0.954	2	8	418.36	N.A.	-	-
87	10950444	0.954	3	7	451.46	3.1	yes	-
88	17892707	0.953	1	2	302.45	4.8	yes	-
89	4620	0.953	1	2	302.45	4.1	yes	36592 443947 20056632
90	12137520	0.952	2	3	242.27	3.4	yes	-
91	229456	0.950	0	3	400.59	7.2	no	3701882
92	20317323	0.949	1	4	358.47	2.7	yes	-
93	21863683	0.948	1	4	472.7	7.1	no	-
94	10007408	0.948	2	8	538.28	3.4	no	16110566 16110573
95	21485696	0.948	1	2	316.48	4.9	yes	-
96	127848	0.947	2	4	346.46	3	yes	14586234 21120905
97	10098582	0.947	3	4	533.54	6.7	no	10368254 10413229 11756875
98	11718642	0.947	0	2	416.64	5.9	no	-
99	21305747	0.946	1	3	446.71	8.5	no	-
100	3008386	0.945	2	5	247.21	1.9	yes	-
101	10716828	0.944	1	3	416.64	6.5	no	-
102	22812421	0.943	1	4	522.76	8.4	no	23338813
103	255263	0.941	1	2	288.42	3.8	yes	-
104	10670683	0.941	1	3	438.64	6.3	no	-
105	10296356	0.940	0	3	430.66	7.4	no	-
106	10600919	0.940	1	4	484.71	7.3	no	-
107	539337	0.940	1	2	414.66	6.8	no	22213015 22213016
108	17978289	0.940	1	6	360.41	5.2	no	-
109	10553031	0.939	3	6	488.66	3.9	yes	-
110	418220	0.939	2	4	574.83	5.9	no	21144389
111	10096634	0.938	1	2	478.75	9.4	no	-
112	21310283	0.938	1	2	424.66	8.2	no	23496855
113	155546	0.938	1	11	1414.11	N.A.	no	-
114	21711898	0.936	0	3	414.62	6.4	no	23617728
115	13328	0.936	0	3	346.5	5.5	no	526981 6428252 6428260 6428269 6428278 11871552 11871553 11871554 11871555 13267937 13267945
116	21485488	0.934	1	2	302.45	4.3	yes	-
117	10254332	0.934	3	4	474.72	5.5	no	10254331 23427152
118	23322678	0.932	2	3	360.53	4	yes	-
119	20301682	0.932	2	3	332.48	2.8	yes	-
120	10390797	0.931	2	5	488.7	5.4	no	10874616 21776462
121	14284356	0.931	1	4	488.74	8.5	no	-
122	23518157	0.930	1	4	460.69	8.2	no	-
123	2748185	0.930	1	3	438.64	7.9	no	2751558 11868007 11868008 11868009 11868010 11868508 11868509 11868510 11868511
124	10006146	0.929	3	6	499.64	5.8	no	10051671
125	10862966	0.926	2	4	432.64	5.7	no	-
126	11554687	0.926	3	4	474.72	5.9	no	-
127	19357239	0.925	2	4	363.53	4.3	yes	-
128	21576162	0.925	2	5	476.69	6.1	no	21775779

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
129	21768874	0.924	1	3	444.69	6.6	no	-
130	20372533	0.921	1	3	400.59	6	no	22801487
131	476483	0.920	1	3	430.66	7.2	no	-
132	543647	0.919	1	3	430.66	6.5	no	-
133	630170	0.918	0	2	400.64	6.8	no	22296271 22296272
134	5284175	0.917	3	5	404.54	4.4	yes	-
135	10624217	0.917	2	5	470.65	8.4	no	-
136	150982	0.915	2	3	304.42	2.4	yes	247021 3496462 10063643 10335203
137	5207588	0.913	2	3	304.42	3.1	yes	14311666 14311667 22842098
138	10424336	0.910	2	2	290.44	4.1	yes	-
139	11636139	0.909	1	3	683.14	15.1	no	-
140	10160599	0.909	0	3	442.67	7.5	no	-
141	19609537	0.904	3	6	624.87	N.A.	no	-
142	526976	0.902	0	3	444.69	8	no	6428246 6428255 6428263 6428272 10275324 20837713 20837719
143	21485770	0.901	1	2	316.48	4.8	yes	-
144	19956936	0.901	1	3	484.41	5.5	no	-
145	538828	0.900	1	4	564.84	9.6	no	-
146	23018386	0.899	1	4	432.64	7.5	no	-
147	18998235	0.898	1	3	390.58	4.9	yes	-
148	10002700	0.897	1	2	428.69	7.1	no	10862890
149	622810	0.896	1	2	414.66	7.3	no	22295493 22295494
150	18186234	0.896	0	3	408.57	6.8	no	-
151	9953948	0.891	2	8	421.4	N.A.	-	-
152	18979959	0.888	1	2	336.53	4.7	yes	-
153	9934361	0.887	2	9	469.46	3.5	yes	-
154	538065	0.887	0	3	346.5	5.5	no	155547 22212857 22212858
155	10225400	0.887	0	3	442.67	7.6	no	-
156	20545999	0.887	2	3	318.45	2.4	yes	-
157	21764096	0.881	1	2	492.78	9.8	no	-
158	10886169	0.880	1	4	553.61	7.4	no	11006178
159	11271103	0.879	2	2	458.76	9.3	no	11442516
160	5090736	0.878	0	3	346.5	5.5	no	-
161	23203405	0.878	3	5	483.64	6.1	no	-
162	20848953	0.877	2	3	304.42	3.5	yes	21252241 21252246
163	14678527	0.877	1	4	404.58	6.1	no	-
164	20372620	0.877	1	2	358.56	6.5	no	22801500
165	11854968	0.876	1	3	306.44	4.9	yes	-
166	10884228	0.876	0	4	416.59	6.6	no	-
167	10275325	0.875	0	3	444.69	7.7	no	-
168	19917239	0.874	1	4	427.37	5.7	no	21120971 22844356 22848245
169	18664606	0.873	5	7	522.71	5.5	no	19063624
170	23599966	0.873	0	2	402.65	8.7	no	-
171	23203398	0.868	3	5	421.57	4.6	yes	-
172	18595214	0.868	1	2	302.45	3.9	yes	21584518
173	10181894	0.867	0	3	442.67	7.5	no	-
174	11619842	0.864	2	4	474.72	6.8	no	-
175	165246	0.861	1	2	288.42	3.5	yes	21632719 21632720 21632721 21632722
176	23504659	0.860	3	4	432.64	4.4	yes	-
177	2752859	0.860	1	4	460.69	7.2	no	3903702 10411948 11887409
178	21802887	0.856	3	6	655.99	9.2	no	-
179	446211	0.855	2	2	346.55	5.3	no	-
180	16110574	0.855	2	7	471.3	2.4	yes	16110570
181	20301550	0.854	2	3	318.45	2.3	yes	-
182	21119280	0.852	2	4	472.7	N.A.	-	-
183	5283975	0.848	2	3	374.56	5.8	no	-
184	463506	0.846	2	5	640.98	10.7	no	-
185	11317146	0.846	1	4	460.69	8.2	no	-
186	9954749	0.845	2	7	437.46	4.8	yes	10717833
187	11350922	0.844	2	2	430.71	8.4	no	11464595
188	5039651	0.844	1	3	436.63	7.7	no	5704602 11886946 11886947 11886948 11886949
189	14862928	0.843	1	5	432.59	4.3	yes	-
190	21776018	0.843	2	3	465.11	6.5	no	-
191	20657237	0.842	2	7	689.99	5.7	no	-
192	3357227	0.841	1	5	494.66	7.8	no	17369984
193	18654558	0.840	1	5	350.45	4.1	yes	19004062
194	5234995	0.839	1	4	558.86	9.9	no	-
195	21776019	0.835	2	3	477.12	6.4	no	-

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
196	23256115	0.835	1	4	539.59	8	no	-
197	21305748	0.834	1	3	430.66	7.6	no	-
198	11198799	0.833	0	3	346.5	6.6	no	-
199	21800712	0.833	1	2	444.73	9.2	no	-
200	1806771	0.832	2	4	308.74	2.4	yes	1806773 1806774 1806775 3755943 7092629 7092630 7092631 7092632 16401523
201	23399602	0.830	1	5	502.73	6.8	no	-
202	2751561	0.830	1	2	394.59	6.7	no	11868522 11868523 11868524 11868525
203	244804	0.829	0	3	360.53	6.1	no	4553983
204	10116956	0.825	0	3	484.75	8.9	no	-
205	582786	0.824	1	4	402.57	5.7	no	-
206	23496884	0.822	1	3	462.66	8.1	no	-
207	17896831	0.821	3	3	418.65	6.8	no	-
208	1783687	0.815	0	3	388.58	7	no	3730220 11876284
209	4464238	0.814	1	4	558.86	10	no	-
210	21802896	0.814	3	6	641.96	8.8	no	-
211	23731380	0.812	1	3	360.53	4.8	yes	-
212	5053709	0.810	1	2	414.66	7.3	no	-
213	10524467	0.809	1	3	390.6	5.4	no	-
214	11178276	0.808	2	2	444.73	8.9	no	11316743
215	20464518	0.808	3	4	643.03	12.6	no	-
216	23201396	0.805	1	4	655.05	14.3	no	-
217	10298155	0.803	0	3	458.72	8.4	no	18186231
218	17946164	0.802	4	6	436.58	3.2	yes	-
219	10275323	0.802	0	3	444.69	7.9	no	-
220	22729961	0.799	2	3	318.45	3.1	yes	-
221	222813	0.797	2	3	346.5	3.4	yes	-
222	21471108	0.796	0	3	366.92	5.5	no	-
223	1843	0.796	0	3	411.37	5.8	no	128336
224	5064819	0.796	2	3	318.45	2.9	yes	-
225	22239603	0.795	1	3	458.72	8.5	no	-
226	21329128	0.795	2	3	304.42	2.2	yes	-
227	2754171	0.794	1	2	382.58	7.1	no	5108126 11868912 11868913 11868914 11868915
228	22900388	0.790	2	10	951.36	14.2	no	-
229	10139348	0.790	0	3	458.72	8.3	no	23343600
230	1759261	0.789	1	5	390.51	4.6	yes	3704729 10093105 11875404 11875406 11875408 11875410 16399798
231	164705	0.787	2	4	332.43	2.7	yes	12849401 12849411
232	11387131	0.785	1	4	539.59	8.2	no	-
233	20564149	0.784	0	2	358.56	6.3	no	-
234	23447000	0.783	2	3	338.87	5.6	no	-
235	20657239	0.782	2	7	689.99	5.9	no	-
236	11212445	0.780	0	3	425.4	6.1	no	-
237	20444491	0.779	3	5	473.69	6	no	22815029
238	21672826	0.777	1	5	825.25	16	no	-
239	22239539	0.775	1	3	432.68	8	no	-
240	16750073	0.775	1	5	446.57	6.3	no	-
241	10628564	0.773	1	3	709.18	16	no	-
242	613004	0.770	1	4	378.55	4.6	yes	-
243	10009206	0.768	2	5	631.68	8.6	no	-
244	11091296	0.766	2	2	430.71	8.3	no	-
245	272885	0.765	2	6	476.65	6.7	no	-
246	236416	0.763	0	3	360.53	6	no	5130359
247	259889	0.763	2	3	332.48	2.6	yes	4597476 10172078
248	32891	0.762	0	3	414.62	6.5	no	-
249	10368167	0.758	1	3	486.77	9.5	no	-
250	11297041	0.752	1	3	709.18	15.5	no	-
251	20638756	0.747	1	7	528.69	8.1	no	22114720
252	10226272	0.746	0	3	456.7	7.9	no	-
253	20301542	0.744	2	3	332.48	2.8	yes	-
254	21485543	0.744	1	2	364.52	5.2	no	-
255	14678529	0.744	1	4	418.61	6.6	no	-
256	20539921	0.742	0	3	456.7	7.9	no	22811518 22811521 22811522
257	20657193	0.733	1	4	506.78	8.4	no	-
258	22090298	0.733	1	4	516.8	10.2	no	-
259	5146437	0.732	2	4	332.43	2.8	yes	-
260	467408	0.731	1	6	676.96	11	no	-
261	20317304	0.728	2	4	374.51	3.7	yes	-
262	11079680	0.728	1	3	390.6	7.5	no	23428088

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
263	10907109	0.725	2	7	490.63	4.6	yes	10939915
264	9930574	0.724	3	7	395.4	2.8	yes	10715687 11794987
265	20056349	0.723	2	5	448.64	5.5	no	-
266	624311	0.722	1	4	502.77	8.1	no	22295675 22295676
267	9808339	0.721	3	5	578.82	N.A.	no	-
268	543443	0.720	1	3	362.55	6	no	22213459
269	4086221	0.719	2	6	412.54	2.5	yes	-
270	10093616	0.717	1	3	420.58	7.2	no	-
271	5496312	0.716	2	5	504.74	6.3	no	6320102 10601630
272	4220520	0.711	1	3	348.52	6.2	no	11840032 11895638 11895639 11895640 11895641
273	22052718	0.708	2	4	460.69	5.7	no	-
274	10297174	0.707	0	3	442.67	7.4	no	-
275	14419	0.705	0	3	332.48	5.2	no	224447 252015 932815 1789782 6428248 6428265 6428274 11809740 11867710 11867711 11867712 21139819 21139820
276	579172	0.699	0	3	332.48	5.2	no	11859745 11890332 11890333 11890334 22216252
277	536358	0.697	0	3	332.48	4.9	yes	22212599
278	269124	0.694	0	3	318.45	4.5	yes	-
279	10504916	0.693	2	2	480.76	10.4	no	10672390 21582717 21582718
280	20368428	0.690	1	2	330.5	5	no	-
281	11262694	0.689	1	3	737.23	16.5	no	-
282	20301524	0.689	2	3	318.45	2.3	yes	22797425
283	23426889	0.687	2	5	490.72	6.4	no	-
284	21711979	0.686	2	5	448.64	6.1	no	-
285	10863978	0.682	1	2	497.22	9.1	no	11733761
286	15560257	0.677	0	0	204.35	3.8	yes	21669871 23425493 23425495
287	21271937	0.672	1	3	466.7	8.6	no	-
288	10253735	0.669	3	3	460.73	7.7	no	-
289	11069762	0.666	2	2	416.68	7.8	no	-
290	25440	0.665	2	2	276.41	3.5	yes	9814176 9835303 9838481 9921701 9965411 10333689 11208152 16760418
291	48619	0.662	0	3	344.49	5.5	no	-
292	3798204	0.661	1	3	536.49	8.8	no	16396085
293	23256112	0.660	1	4	446.66	6.8	no	-
294	1782124	0.658	0	3	360.53	5.9	no	3814440 11876243
295	2738135	0.657	3	4	390.56	4.4	yes	5233394 5283976 11887975 11887977 11887979 11887981
296	9982734	0.656	1	6	482.64	7.5	no	-
297	10054282	0.654	2	5	598.9	9.1	no	-
298	715670	0.648	1	4	265.66	3.4	yes	-
299	21312299	0.648	1	4	310.11	3.5	yes	21312290
300	23203385	0.638	3	5	435.6	5.2	no	-
301	20841928	0.637	1	3	416.62	4.7	yes	-
302	23322687	0.635	2	3	360.53	4.2	yes	-
303	13132651	0.622	1	4	404.58	5.5	no	13132650
304	10160743	0.620	0	3	444.69	7.9	no	-
305	13987	0.619	0	3	394.55	7	no	251107 6994432 10069228 11862170 11867725 11867726 11867727 11867728 11872176 11872177 11872178 16213018 16395308 17376414
306	22767171	0.616	3	4	376.53	2.3	yes	-
307	20669538	0.614	1	4	510.75	6.8	no	-
308	22212600	0.611	0	3	332.48	4.9	yes	-
309	10393492	0.608	1	3	584.61	9.3	no	-
310	23931165	0.601	1	2	274.4	3.2	yes	-
311	107497	0.599	0	3	346.5	5.7	no	3530540 10926061
312	21420874	0.599	1	2	344.53	5.5	no	22798798
313	10098401	0.598	2	2	526.86	10.2	no	-
314	526977	0.587	0	3	472.74	8.9	no	6428247 6428256 6428264 6428273 10205571 20837714 20837720
315	11225239	0.584	3	4	472.7	6.8	no	11442846
316	20465753	0.577	3	4	643.03	12.6	no	-
317	10595721	0.575	1	3	376.57	5.1	no	-
318	10715264	0.574	2	5	387.51	3.5	yes	-
319	43373	0.574	1	2	330.5	4.9	yes	4068828 9797605 16627138 22798805
320	15981461	0.567	1	4	390.56	5.6	no	-
321	23203432	0.565	3	5	435.6	5.1	no	-
322	18974326	0.562	0	3	406.56	5.7	no	23615854
323	192428	0.557	1	6	535.72	N.A.	no	-
324	521890	0.557	2	4	346.46	1.4	yes	222804
325	22239584	0.556	1	3	444.69	8.2	no	-
326	250295	0.554	0	2	276.41	6	no	3249341
327	10551934	0.551	1	4	460.69	7	no	-
328	5113668	0.550	1	1	289.43	N.A.	-	-
329	13314933	0.548	0	4	478.64	3.5	yes	13314935

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
330	11834066	0.543	1	2	316.48	4.5	yes	-
331	22898971	0.539	2	8	414.38	N.A.	-	-
332	10040730	0.534	2	3	304.42	2.3	yes	-
333	10027889	0.529	2	4	474.72	6.2	no	11684318
334	18472968	0.528	1	2	316.48	4.5	yes	-
335	23203414	0.520	3	5	421.57	4.7	yes	-
336	20693842	0.511	0	4	488.74	8.7	no	-
337	21344437	0.510	1	4	460.69	7.7	no	21863833
338	457928	0.508	2	6	584.83	8.1	no	11505221 20111792 22902865
339	22166383	0.507	1	5	532.79	N.A.	no	-
340	10766245	0.506	1	3	456.7	7.5	no	-
341	10645045	0.500	1	4	406.6	6.9	no	-
342	11983304	0.500	1	4	404.58	6.2	no	-

1: PubChem Compound ID, 2: predicted probability, 3: number of H bond donor, 4: number of H bond acceptor, 5: molecular weight, 6: whether a chemical compound satisfies Lipinski's rule of five (# of H bond donor $\leq$ 5 & # of H bond acceptor $\leq$ 10 & molecular weight $\leq$ 500 & xlogp $\leq$ 5), 7: not available because some values were not available.

Supplementary Table A43 Details of predictions in Fig. III.16B

	1 ID	2 prob.	3 H don.	4 H accept.	5 M.W.	xlogp	6 Lipinski5	compounds of same connectivity
1	4086221	1.000	2	6	412.54	2.5	yes	-
2	17892707	1.000	1	2	302.45	4.8	yes	-
3	11687	0.999	2	4	346.46	2.6	yes	627480 22296005
4	21485488	0.999	1	2	302.45	4.3	yes	-
5	125155	0.999	2	3	416.64	5.4	no	21120484
6	630661	0.999	2	4	336.44	2	yes	-
7	623740	0.999	2	3	304.42	1.8	yes	150968 1149869 1778898 1778899 1778901 6546023 11872739 11872740 16395893 22849351
8	19609992	0.999	2	3	304.42	2.6	yes	-
9	23724679	0.998	2	2	418.7	8.1	no	-
10	14259153	0.998	2	3	304.42	1.9	yes	23256526
11	19027438	0.998	3	5	362.46	1.4	yes	-
12	164705	0.998	2	4	332.43	2.7	yes	12849367
13	3733299	0.997	2	4	329.43	2.5	yes	16396395
14	65545	0.997	2	3	304.42	3	yes	237641 9904576 10086383 15708880 21139669 23624176
15	21709398	0.997	2	3	318.45	2.6	yes	-
16	23037249	0.997	1	4	415.59	7 N.A.	-	-
17	21485696	0.997	1	2	316.48	4.9	yes	-
18	130408	0.996	2	4	346.46	1.8	yes	21121280
19	20713669	0.995	1	2	274.4	3.7	yes	-
20	21672107	0.994	2	3	414.62	4.8	yes	-
21	23322678	0.993	2	3	360.53	4	yes	-
22	13314930	0.993	1	3	354.48	3	yes	13314933
23	20743369	0.993	2	4	441.65	4.9	yes	23504634
24	14324907	0.991	2	4	819.42	22.3	no	-
25	11611466	0.990	1	7	419.4	2.8	yes	-
26	22958114	0.989	1	2	330.5	4.6	yes	-
27	107835	0.989	2	3	332.48	3.7	yes	-
28	11539439	0.988	2	7	439.86	N.A.	-	-
29	150854	0.988	1	2	288.42	3.8	yes	10039774 10085744 21122962
30	482758	0.987	2	5	542.79	7.4	no	21147722
31	11698085	0.987	1	7	449.49	4	yes	-
32	18753323	0.986	1	8	445.44	3.1	yes	-

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
33	22767171	0.986	3	4	376.53	2.3	yes	-
34	251593	0.986	2	3	304.42	2.9	yes	4038155
35	11539441	0.985	2	7	439.86	N.A.	-	-
36	521890	0.985	2	4	346.46	1.4	yes	222804
37	11575793	0.984	2	7	439.86	N.A.	-	18753348
38	439534	0.983	2	4	320.42	2.4	yes	-
39	21329128	0.983	2	3	304.42	2.2	yes	-
40	4435	0.981	0	3	406.56	5.6	no	229455 11869560 11872989 16394511 16396215 23615856
41	11532211	0.980	1	7	437.85	3.9	yes	-
42	21119280	0.980	2	4	472.7	N.A.	-	-
43	248649	0.980	2	3	304.42	3.5	yes	3994025
44	18753367	0.980	1	8	421.39	3.4	yes	-
45	11719006	0.979	1	9	433.44	2.4	yes	-
46	18753324	0.979	1	7	437.85	3.9	yes	-
47	9826529	0.979	1	7	482.3	4.1	yes	18753320
48	11539442	0.979	1	7	403.4	3.2	yes	-
49	20301682	0.979	2	3	332.48	2.8	yes	-
50	11539440	0.978	1	7	403.4	3.9	yes	-
51	9802277	0.977	1	8	421.39	3.4	yes	18753335
52	19609537	0.976	3	6	624.87	N.A.	no	-
53	11163336	0.976	1	2	330.5	5	no	11198308 11370933 11739019
54	11575794	0.975	1	7	403.4	3.2	yes	18753349
55	247863	0.974	1	2	304.47	4.6	yes	244128
56	11163334	0.974	1	2	330.5	5	no	11186567 11290455 11484329
57	22868519	0.974	1	2	318.49	5.8	no	23358780
58	18637525	0.973	1	2	290.44	5	no	18637528 19868612
59	9876052	0.972	2	6	869.31	N.A.	no	-
60	10950444	0.972	3	7	451.46	3.1	yes	-
61	2748117	0.972	2	4	350.49	5.5	no	11867730 11867732 11867734 11867736
62	23496884	0.970	1	3	462.66	8.1	no	-
63	23599899	0.969	1	2	334.54	7.1	no	-
64	18637620	0.969	1	2	304.47	4.6	yes	18978453
65	9973200	0.969	3	3	318.5	4.1	yes	-
66	2375	0.968	2	9	430.37	2.5	yes	56069 441405 16110572
67	536506	0.968	1	2	318.49	5.4	no	102347 23794292
68	20092534	0.968	1	2	318.49	5.2	no	-
69	10005449	0.968	1	7	482.3	4.1	yes	23082061
70	237186	0.967	1	2	318.49	5.2	no	21117221 21117222
71	97787	0.967	2	3	171.19	0.7	yes	-
72	18464404	0.967	2	3	444.69	6.1	no	18651264
73	3397	0.967	1	6	276.21	2.6	yes	10802904
74	3387	0.967	2	4	336.44	2	yes	6446 6429867 6432701 10088392 20056778
75	5881	0.967	1	2	288.42	3	yes	76 134506 719264 908452 1780096 1780097 2817714 3036244 5318142 6432532 6603817 7048561 7092995 9814492 9860744 10085638 10541387 10613370 11889770 12358559 12358563 16401868 16759247 23968375
76	4493	0.967	1	7	317.22	1.8	yes	-
77	2756	0.967	3	6	252.34	1	yes	-
78	10203	0.967	1	2	290.44	4.3	yes	225 5879 5880 11303 247732 441302 736280 908626 1274488 1769120 1769123 2754097 2754130 2825511 5318140 5318590 5701989 6420080 6432531 6432554 6603721 10732277 10780013 11055399 11109055 11243255 11437899 11868646 11868647 11868648 11868649 11868766 11868767 11868768 11868769 11869362 11869363 11869364 11869365 11875890 11890326 11890327 11957461 12306723 12306736 16401171 16758042 16759306 23729586 24131069
79	15	0.967	1	2	290.44	4.5	yes	10635 11302 147320 667503 968803 1779160 1779163 1779165 2748143 6432518 6951009 9965810 10379482 10780012 10902374 11860392 11867857 11867858 11867859 11867860 11869358 11869359 11869360 11869361 11914201 11920866 12446692 12446728 16111632 16395945 23729584

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
80	5833	0.967	0	4	416.57	3.4	yes	5267 941624 1255733 1425524 1742265 1742266 6419955 6432512 6604008 6604182 6710654 10477021 11509929 11863525 11869418 11869419 11869420 11869578 11869579 11869580 11957686 16394537 16757735 16760181 18594034
81	2914	0.967	0	4	416.94	3.2	yes	9880 3034800 5702030 6603774 6714003 9823237 9853228 16757655 16759143 21563805
82	193714	0.967	1	2	164.25	0.9	yes	10001057
83	6128	0.967	0	2	286.41	3	yes	2193 249335 517670 712379 719687 1777653 1777654 1777655 2825552 6432604 6603719 7092858 7092862 10636976 11862468 11869385 11869386 16401731 22211566
84	3170	0.967	0	3	360.53	5.8	no	224004 10316496 16627147 20056767 23615700
85	10634	0.967	2	2	290.44	3.6	yes	102192 223407 968807 1778952 1778953 9814541 9904098 9922062 9965813 10062826 11872752 11872753 11872754 16395906 18444027
86	5878	0.967	1	3	306.44	4.2	yes	249098 6429885 6432530 9883009 23616084
87	10204	0.967	1	2	288.42	3.6	yes	5408 6013 186662 242356 369323 638016 711503 908606 908607 908608 1312122 2825665 3034651 5701998 6432567 6541847 7048589 10063008 10108190 10469710 11514789 11861311 11862155 11869421 11869422 11869423 11870497 11890585 11890586 11920270 12304532 12304539 12304540 16220011 16758160 19615208 20848917 22876164
88	21532454	0.967	1	3	299.24	N.A.	-	-
89	23495	0.967	2	3	352.9	2.1	yes	235091 3427265
90	21420874	0.966	1	2	344.53	5.5	no	22798798
91	629113	0.966	1	2	318.49	5	no	21634213 22296179
92	9802031	0.966	2	2	416.68	7.7	no	-
93	10540444	0.965	1	2	276.41	4.1	yes	10826346 10850094 10850095
94	18753361	0.965	1	10	471.4	4.2	yes	-
95	11209293	0.965	1	2	316.48	5.4	no	-
96	10507285	0.965	1	2	570.89	12.5	no	10698293 11801190 21582716
97	10297173	0.964	0	3	442.67	7.3	no	-
98	5039651	0.964	1	3	436.63	7.7	no	5704601 11886946 11886947 11886948 11886949
99	6011	0.964	1	2	304.47	4.8	yes	3167 12446690 16627148 22794328 22821231 23615698 23615699
100	10087279	0.964	1	2	318.49	6.2	no	-
101	4157657	0.963	1	3	315.45	3.7	yes	-
102	23599901	0.963	1	2	320.51	6.8	no	-
103	25012	0.961	1	2	304.47	4.8	yes	240037 11954137 22790919
104	10017941	0.961	1	2	304.47	5.8	no	-
105	22026903	0.961	0	1	504.83	12.6	no	-
106	3044053	0.959	1	2	370.92	N.A.	-	-
107	3834333	0.958	1	5	450.57	4.4	yes	16403114 24023724
108	23322692	0.958	1	4	374.51	4.1	yes	-
109	5233916	0.958	1	2	318.49	4.8	yes	-
110	134100	0.958	1	2	416.34	5.3	no	-
111	20372620	0.957	1	2	358.56	6.5	no	22801500
112	449706	0.957	1	3	307.43	4.6	yes	451177 9972569 10448011
113	23322702	0.957	1	2	342.51	5.4	no	-
114	9800381	0.956	2	3	387.35	N.A.	-	-
115	3084142	0.956	0	3	416.64	6.8	no	13014290 13014291
116	632284	0.956	1	2	360.57	5.8	no	22296515 22296516
117	21059273	0.956	1	2	316.48	4.9	yes	-
118	21576162	0.955	2	5	476.69	6.1	no	21775779
119	23599906	0.955	1	2	348.56	7.7	no	-
120	19793642	0.954	1	2	306.48	6.3	no	24124754
121	13917178	0.953	2	3	294.43	5.6	no	18638064
122	10841402	0.953	4	7	608.8	8.3	no	-
123	24043345	0.953	1	3	315.45	3.1	yes	-
124	15020	0.952	1	2	304.47	5	no	451146 616642 10590720 10709582 11869651 11869652 11869653 16394601 22295049 22809684
125	10001994	0.951	2	2	416.68	8	no	-

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
126	23203414	0.951	3	5	421.57	4.7	yes	-
127	21485603	0.950	0	3	434.61	6.8	no	22790420
128	21327828	0.950	2	4	418.61	5.7	no	-
129	386134	0.949	1	4	420.61	6.2	no	4601615
130	12137491	0.948	0	7	357.33	6	no	-
131	2748151	0.948	1	2	388.54	5.7	no	4272506 11886938 11886939 11886940 11886941
132	21610139	0.948	2	4	512.74	6.5	no	-
133	9807705	0.947	3	5	553.77	6.6	no	-
134	20758452	0.945	2	6	446.58	4.4	yes	-
135	418220	0.945	2	4	574.83	5.9	no	21144389
136	15227104	0.945	1	2	369.34	4.9	yes	20564133
137	21436611	0.945	1	2	324.89	4.6	yes	22797535
138	15444	0.944	1	3	308.43	4.4	yes	242489 519277 21139712
139	248277	0.944	1	2	362.93	5.6	no	5105189
140	10091357	0.943	1	2	383.36	5.5	no	-
141	23599903	0.943	1	2	320.51	6.7	no	-
142	23677211	0.943	1	4	412.54	N.A.	-	-
143	23365273	0.942	4	7	486.69	5	no	-
144	23931165	0.940	1	2	274.4	3.2	yes	-
145	20564132	0.940	1	3	322.46	4.8	yes	-
146	10007408	0.940	2	8	538.28	3.4	no	16110566 16110573
147	10271294	0.938	1	2	383.36	5.3	no	11246054
148	172080	0.938	2	5	362.46	1.7	yes	21125036
149	13328	0.938	0	3	346.5	5.5	no	526981 6428252 6428260 6428269 6428278 11871552 11871553 11871554 11871555 13267933 13267935
150	20564162	0.938	1	2	304.47	4.8	yes	-
151	2748147	0.937	1	2	369.34	4.7	yes	4044257 11867876 11867877 11867878 11867879
152	240038	0.937	1	2	318.49	5.1	no	21139694 21139695
153	273381	0.936	1	2	304.47	5.1	no	-
154	21019537	0.935	3	8	640.85	7.9	no	-
155	21558457	0.935	3	3	348.52	4.6	yes	-
156	21485576	0.935	1	2	344.53	5.7	no	-
157	22833522	0.934	2	6	384.49	2.6	yes	-
158	11662272	0.933	1	7	445.45	1.6	yes	-
159	18988333	0.933	3	6	634.88	N.A.	no	-
160	482757	0.932	4	6	753.83	17.2	no	21147721
161	627591	0.932	1	2	304.47	4.4	yes	21634211
162	18651628	0.931	1	5	336.42	3.6	yes	19004061
163	20847746	0.931	3	3	430.66	6.5	no	-
164	10529114	0.930	4	5	490.72	4.3	yes	-
165	9570509	0.930	1	5	410.51	5.5	no	20843870
166	165246	0.929	1	2	288.42	3.5	yes	21632719 21632720 21632721 21632722
167	11529815	0.929	1	3	318.45	2.6	yes	-
168	13833955	0.929	2	4	346.46	2.1	yes	18595696
169	11677799	0.929	2	5	504.74	6.9	no	-
170	22239626	0.928	1	5	490.72	7.7	no	-
171	10674597	0.926	1	2	570.89	12.6	no	10769399
172	21389525	0.923	2	5	467.62	N.A.	-	-
173	18396884	0.922	1	2	470.43	6	no	19365739
174	20471506	0.922	4	7	650.88	N.A.	no	-
175	10715264	0.921	2	5	387.51	3.5	yes	-
176	19977850	0.921	4	7	474.68	5.1	no	-
177	15607478	0.921	2	4	360.49	3	yes	-
178	10093616	0.920	1	3	420.58	7.2	no	-
179	11574935	0.920	1	2	402.65	7.8	no	11704092
180	20547436	0.920	1	2	318.49	5.8	no	22819626
181	22767169	0.920	2	6	432.55	2.8	yes	-
182	9995441	0.920	1	3	308.43	3.8	yes	10267286
183	2748175	0.919	1	2	288.42	3.1	yes	3639419 11886958 11886959 11886960 11886961
184	20657372	0.917	1	5	490.72	7.9	no	-
185	20847658	0.915	3	3	444.69	6.3	no	-
186	21632969	0.913	3	4	474.72	5.5	no	-
187	259866	0.913	2	4	376.53	4.2	yes	-
188	10812094	0.912	2	5	420.58	5.1	no	-
189	11167846	0.912	1	2	494.79	10	no	-
190	23322668	0.911	1	2	322.87	4.5	yes	-

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
191	10245130	0.911	1	2	314.46	4.1	yes	-
192	351790	0.910	0	3	330.46	4.7	yes	-
193	313275	0.910	0	3	414.62	6.9	no	101692 160498 5320116 5321141 5321359 6452468 6708616 10916686 12304423 12304429 21140634 21140635
194	10926115	0.910	1	4	348.48	4.7	yes	-
195	451184	0.909	1	3	307.43	3.8	yes	130696 9972570 10040953 10357878 21121324
196	2754109	0.908	1	4	376.53	5.7	no	4316021 11868675 11868677 11868679 11868681
197	11364869	0.907	1	4	550.81	10.2	no	-
198	10915796	0.907	1	4	376.53	4.5	yes	-
199	9931282	0.906	2	6	408.44	3.5	yes	18753351
200	21800712	0.906	1	2	444.73	9.2	no	-
201	19603132	0.905	2	8	418.36	N.A.	-	-
202	11530594	0.905	1	1	360.62	8.1	no	-
203	242500	0.903	2	4	346.46	1.1	yes	3957732
204	23365227	0.902	4	7	472.66	4.6	yes	-
205	250299	0.901	1	3	396.56	7.4	no	2748115 11867717 11867718 11867719 11867720 11872195 11872196 11872197 11872198 16395322 23768374 23945618
206	21435914	0.901	2	3	304.42	3.4	yes	22797419
207	10739773	0.901	2	4	400.55	4.5	yes	-
208	17814850	0.897	2	8	418.42	3.1	yes	-
209	21580208	0.896	1	5	502.73	6.8	no	-
210	21917762	0.896	1	4	423.56	N.A.	-	-
211	9966551	0.894	1	2	312.45	5.2	no	19861705
212	16627134	0.894	1	2	360.57	6.7	no	-
213	10596088	0.894	1	3	382.54	6.8	no	-
214	10341837	0.894	1	2	416.34	4.5	yes	-
215	20344915	0.894	1	5	362.46	2.7	yes	22810486
216	5272689	0.892	2	3	444.69	7.2	no	15488951 15488952
217	23504659	0.891	3	4	432.64	4.4	yes	-
218	10542856	0.891	1	3	308.43	3.8	yes	-
219	197144	0.890	2	3	312.43	3.7	yes	2748144 4555496 11867861 11867862 11867863 11867864
220	16731463	0.889	2	3	440.66	5.9	no	-
221	9795964	0.888	1	2	288.42	4.2	yes	22495313
222	5283997	0.888	2	4	386.52	4.1	yes	-
223	21511858	0.885	1	3	336.51	N.A.	-	-
224	287615	0.884	1	2	304.47	4.9	yes	11044933
225	250295	0.883	0	2	276.41	6	no	3249341
226	11124404	0.881	3	5	540.77	7.9	no	-
227	11692717	0.880	2	5	576.81	8.2	no	-
228	21989803	0.879	0	2	445.43	6.5	no	-
229	4307654	0.879	0	4	610.87	11.7	no	11614200
230	18655913	0.879	1	2	302.45	4.1	yes	18959518
231	127848	0.878	2	4	346.46	3	yes	14586206 21120905
232	2825516	0.878	0	3	332.48	4.6	yes	11890347 11890348 11890349 11890350
233	10391043	0.876	3	4	494.71	8.1	no	21781221
234	10531074	0.876	0	2	568.87	12.3	no	-
235	561252	0.873	1	5	572.84	8.9	no	22212053
236	301438	0.873	1	2	332.52	5.7	no	-
237	21333139	0.872	2	3	332.48	2.4	yes	-
238	534502	0.870	2	4	360.49	2.1	yes	22212449
239	23599967	0.870	1	3	404.63	7.4	no	-
240	4073988	0.870	1	4	362.5	5.9	no	20975220 23857578
241	21852715	0.869	0	1	352.55	8.2	no	-
242	314376	0.867	2	3	418.65	5.8	no	-
243	222791	0.867	2	3	290.4	1.6	yes	3827295
244	10745295	0.867	1	5	552.78	8.5	no	-
245	19977810	0.866	4	5	446.67	5.1	no	-
246	23365204	0.866	3	5	444.65	5.2	no	-
247	23427316	0.866	2	5	376.49	4.9	yes	-
248	10054590	0.865	3	6	618.84	8.6	no	11758552
249	21333095	0.865	2	3	346.5	2.9	yes	-
250	18599333	0.863	2	4	386.52	4	yes	20739520 23509324
251	17991413	0.862	1	2	304.47	4.8	yes	-
252	244912	0.861	1	3	292.41	3.5	yes	4222325
253	14537010	0.858	2	4	334.45	1.4	yes	-
254	4620	0.858	1	2	302.45	4.1	yes	36592 443947 20056632
255	18653138	0.857	1	2	352.94	5.1	no	-

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
256	228492	0.857	1	2	397.39	5.4	no	-
257	10340243	0.857	1	3	388.58	7	no	-
258	463524	0.855	4	5	655.99	9.5	no	10818072 21146497
259	11449978	0.855	1	2	318.49	5.8	no	-
260	235676	0.855	1	2	304.47	5	no	5165408 22790921 22790931
261	20301538	0.854	3	3	334.49	3.2	yes	-
262	11339469	0.852	1	4	432.62	5.9	no	-
263	114243	0.851	2	4	360.49	3.5	yes	17872349 21119207 22791151
264	3008386	0.851	2	5	247.21	1.9	yes	-
265	397618	0.850	2	8	700.98	10.4	no	500414 22210658
266	18635375	0.850	1	2	288.42	3.2	yes	19845446
267	10790705	0.848	1	5	474.65	5	no	-
268	21522611	0.848	2	4	440.61	5	no	-
269	19774473	0.847	2	4	346.46	1.2	yes	-
270	5207588	0.847	2	3	304.42	3.1	yes	14311634 14311635 22842098
271	10257971	0.842	1	7	597.78	9	no	-
272	10145526	0.841	1	5	606.87	9.5	no	23547770
273	526977	0.841	0	3	472.74	8.9	no	6428247 6428256 6428264 6428273 10205571 20837714 20837720
274	17882638	0.840	1	2	288.42	3.7	yes	-
275	11539737	0.839	1	7	417.43	3.4	yes	-
276	22854578	0.838	1	3	398.58	5.5	no	23356172
277	20531740	0.838	2	3	332.48	3.8	yes	22820966
278	11633120	0.837	2	7	419.4	2.6	yes	-
279	14351374	0.835	1	5	473.69	5	no	-
280	23447021	0.835	1	2	304.47	4.7	yes	-
281	10690961	0.835	2	3	376.57	4.6	yes	-
282	18718335	0.833	3	6	614.9	8.8	no	22620443
283	5316088	0.833	2	4	360.49	2.2	yes	-
284	21711932	0.833	1	5	459.62	N.A.	-	-
285	414397	0.833	2	4	346.46	3.1	yes	24187067
286	19896496	0.831	0	2	458.67	8.6	no	22844093
287	21435916	0.831	2	3	318.45	3.6	yes	-
288	21578406	0.830	1	2	383.36	5.2	no	-
289	449649	0.830	1	2	428.36	4.9	yes	10094199
290	261357	0.828	1	5	558.81	8.7	no	-
291	21711954	0.827	1	2	312.45	3.5	yes	23616850
292	3577208	0.827	3	4	445.62	3.2	yes	10527325 16399356
293	287616	0.827	1	2	290.44	4.7	yes	-
294	11529598	0.826	1	2	304.47	4.8	yes	-
295	22214042	0.826	3	3	403.39	3.5	yes	-
296	19386002	0.825	1	3	464.68	7.5	no	22856060
297	449651	0.823	1	3	321.46	4.2	yes	10358784
298	10573358	0.823	1	5	405.53	5.1	no	23365183
299	10005226	0.822	2	3	476.77	8.9	no	10027981 10028102 10367854
300	19392397	0.822	1	2	346.55	6.8	no	22855198
301	11209716	0.821	1	2	330.5	4.7	yes	11232791 11313514 11313515
302	23379055	0.820	0	2	358.56	5.8	no	-
303	22151833	0.819	1	6	504.7	6.9	no	-
304	439618	0.817	1	2	290.44	4.9	yes	234464 16218953 20599279 20758703 21139625
305	11668857	0.814	1	6	417.45	6.3	no	-
306	18464409	0.810	1	4	429.61	N.A.	-	-
307	19418208	0.810	1	3	316.43	1.1	yes	-
308	561925	0.809	1	2	318.49	5.1	no	2754136 11868792 11868793 11868794 11868795 14681461 22215019
309	21494714	0.808	1	2	328.49	4.8	yes	-
310	224003	0.806	2	4	350.47	2.3	yes	11954104
311	397622	0.806	2	8	757.09	12.3	no	500418 9875399 21144064 22210660
312	10852075	0.803	1	2	304.47	4.7	yes	11185797 11278126 23374746
313	23681329	0.803	0	7	640.91	N.A.	no	-
314	10698336	0.799	3	5	572.86	8	no	21606390
315	23203408	0.798	3	5	449.62	5.2	no	-
316	11223329	0.798	1	2	400.64	6.6	no	11749960
317	134793	0.798	2	4	552.83	10.6	no	-
318	20657199	0.798	1	5	492.73	8.4	no	-
319	21494700	0.797	1	2	300.44	3.7	yes	-
320	15958446	0.797	3	6	618.84	8.6	no	-

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
321	625566	0.796	3	4	348.48	2.6	yes	107483 160794 9902917 10427858
322	2754188	0.795	0	2	380.56	6.7	no	3936566 11825077 11868964 11868965 11868966 11868967
323	2748165	0.794	1	3	434.61	7.8	no	2751557 11867935 11867936 11867937 11867938 11868504 11868505 11868506 11868507
324	10556204	0.794	0	4	648.98	12.6	no	-
325	21571298	0.790	3	7	676.92	N.A.	no	-
326	21674177	0.790	1	3	428.65	6.2	no	-
327	623875	0.789	1	2	276.41	3.7	yes	128251 9548745 10016318 10446174 10891128 11778226 14009223 14009224 16679839 16679840 22795659
328	10320641	0.788	1	8	433.43	3.2	yes	23082062
329	11766938	0.786	1	6	566.77	8.1	no	-
330	23245684	0.783	3	6	506.71	5.9	no	-
331	4150716	0.782	1	2	352.94	5.1	no	-
332	4318337	0.782	1	2	397.39	5.4	no	-
333	18978374	0.782	1	2	314.46	4.2	yes	22881860
334	23496841	0.779	1	2	382.58	6.9	no	-
335	95345	0.778	2	3	175.01	0.5	yes	5315017 9796565 9818026
336	9824174	0.776	1	8	433.43	3.8	yes	11690589 18753307
337	150982	0.775	2	3	304.42	2.4	yes	247021 3496462 10063643 10335203
338	4071520	0.775	0	3	520.79	10.7	no	-
339	20591736	0.773	1	3	446.71	8	no	-
340	22620461	0.773	4	6	588.86	6.9	no	-
341	23496946	0.773	1	2	396.61	7.3	no	-
342	573668	0.770	1	2	288.42	4.2	yes	-
343	11734400	0.768	0	5	558.83	9.3	no	23246604
344	20657192	0.767	3	7	568.81	6.3	no	-
345	10439604	0.765	5	11	686.89	4.6	no	-
346	17757569	0.762	2	5	364.48	3.5	yes	-
347	9836032	0.761	2	3	302.41	3.3	yes	9922452 9944239 9944240 22726705
348	20545999	0.759	2	3	318.45	2.4	yes	-
349	11417157	0.758	1	3	362.55	6.5	no	-
350	10342550	0.758	3	7	428.57	3.5	yes	-
351	21485685	0.758	2	2	314.46	4.1	yes	-
352	20266777	0.756	2	3	362.55	5.9	no	-
353	19896482	0.753	0	2	416.59	7.1	no	22844089
354	11703468	0.752	2	4	372.5	4.2	yes	-
355	22620460	0.752	3	6	587.85	N.A.	no	-
356	273701	0.752	1	2	312.45	3.5	yes	-
357	2754232	0.752	0	1	366.58	8.2	no	3958712 11868976 11868977 11868978 11868979
358	10746420	0.751	5	8	618.84	5.8	no	-
359	16745166	0.751	1	3	322.46	4.4	yes	16757512
360	10205718	0.750	1	3	474.67	7.3	no	-
361	22937478	0.750	1	3	624.04	N.A.	no	-
362	190884	0.748	8	15	1110.22	N.A.	no	-
363	11145701	0.747	1	1	514.82	12.1	no	-
364	461260	0.745	2	6	586.84	8.9	no	-
365	10572849	0.742	1	4	396.52	5.6	no	-
366	13982078	0.741	0	5	514.74	7.4	no	18634628 18634629
367	21400774	0.740	1	3	344.49	5.1	no	22800005 22800006 22800007
368	3666454	0.740	3	5	429.55	3	yes	10693893 11875301 11875302 11875303 11875304 16399614
369	11762522	0.736	3	5	475.68	6.6	no	-
370	21355874	0.734	2	5	347.43	N.A.	-	22999989
371	250804	0.734	0	3	358.51	5.6	no	250803 3436837
372	10404549	0.734	1	2	332.52	5.6	no	11267583 11359511
373	11340099	0.731	1	3	458.72	7.9	no	21580206
374	2754147	0.730	1	3	332.48	3.7	yes	11868836 11868837 11868838 11868839
375	17882785	0.729	1	2	369.34	4.3	yes	-
376	21393623	0.729	1	2	288.42	3.5	yes	22800529
377	21543037	0.728	1	2	397.39	6.5	no	-
378	235940	0.727	1	2	314.46	4.3	yes	11867638 11867639 11867640 11867641 11954111
379	10623174	0.723	1	4	444.65	6.5	no	21776192
380	17850612	0.722	5	7	604.86	6.3	no	-
381	16664224	0.721	2	4	512.76	7.7	no	18477792
382	21410876	0.720	1	2	276.41	4.6	yes	22795658
383	10168626	0.719	0	4	664.96	11.8	no	-
384	590128	0.719	0	1	462.75	10.7	no	-
385	23504636	0.717	3	5	502.73	5.7	no	-

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
386	18644489	0.715	0	3	329.48	3.9	yes	20065371
387	20657273	0.714	3	6	533.74	4.6	no	-
388	23447000	0.712	2	3	338.87	5.6	no	-
389	23379091	0.712	0	2	358.56	6.1	no	-
390	20669538	0.705	1	4	510.75	6.8	no	-
391	9570750	0.704	1	5	410.51	5.5	no	20843926
392	21118383	0.701	1	4	440.61	8.5	no	22016315 22802754
393	476482	0.701	4	8	266.26	2.8	yes	-
394	10472215	0.695	2	4	336.47	4.6	yes	-
395	21485686	0.695	1	2	330.5	5.1	no	-
396	10698620	0.694	4	6	588.86	7.1	no	15816524 15816525
397	4286359	0.692	1	3	332.48	3.7	yes	-
398	11539736	0.690	2	7	453.89	N.A.	-	-
399	19792827	0.690	1	4	418.61	6.2	no	22847910
400	21958448	0.687	2	3	332.48	5.8	no	23531771
401	2754164	0.684	0	2	378.55	5.8	no	4015910 11888137 11888138 11888139 11888140
402	446188	0.683	2	2	290.44	3.9	yes	136297 164890 519587 9965811 9965812 12476620
403	18336216	0.683	1	3	374.56	6.7	no	-
404	20693838	0.682	3	6	655.95	7.6	no	-
405	2748171	0.680	1	2	328.49	4.9	yes	5131208 11867955 11867956 11867957 11867958
406	255262	0.680	2	4	576.85	N.A.	no	-
407	17376422	0.677	0	2	406.6	8.5	no	19590368
408	22328355	0.671	1	5	420.58	6.1	no	-
409	6822842	0.667	2	3	423.63	6.6	no	11862568 11875801 11875802 11875803 11875804 16400867
410	4419266	0.666	0	3	402.61	6.9	no	11877528 11877529 11877530 11877531
411	20450034	0.666	1	2	260.37	3.3	yes	22792579
412	23599904	0.665	1	2	320.51	6.7	no	-
413	12918583	0.664	1	4	556.84	8.7	no	-
414	23365277	0.663	4	6	468.63	4.6	yes	-
415	23599902	0.663	1	2	320.51	6.8	no	-
416	21800689	0.658	0	2	554.93	11.5	no	-
417	20525532	0.654	3	4	334.45	3	yes	-
418	497950	0.654	1	3	320.47	4.5	yes	11024869
419	22081421	0.653	3	6	252.34	N.A.	-	-
420	11450853	0.649	1	2	346.55	6.2	no	-
421	22620427	0.647	4	7	603.85	N.A.	no	-
422	20268737	0.646	1	4	438.6	N.A.	-	23617454
423	20657242	0.645	2	6	621.91	6.5	no	-
424	11718642	0.642	0	2	416.64	5.9	no	-
425	20693842	0.640	0	4	488.74	8.7	no	-
426	21711974	0.637	1	4	409.54	N.A.	-	-
427	10694229	0.636	0	5	436.62	7.2	no	-
428	10166946	0.632	2	7	574.79	6	no	-
429	622096	0.630	1	5	366.47	3.6	yes	22295444 22295445
430	2748130	0.629	1	2	304.47	4.9	yes	3416437 11867801 11867802 11867803 11867804
431	20693841	0.626	2	4	571.87	8.6	no	-
432	15292632	0.626	3	4	114.11	2.1	yes	23365184
433	11756679	0.624	5	6	527.69	3.3	no	-
434	21693165	0.622	1	3	580.92	11.7	no	-
435	587223	0.621	0	3	414.62	6.9	no	22216786 22216787
436	10504846	0.620	3	7	478.62	4.6	yes	-
437	10721577	0.615	0	6	558.79	8.1	no	-
438	20657252	0.615	0	4	596.92	11.6	no	-
439	16067875	0.613	1	2	211.66	2.3	yes	-
440	20593676	0.612	1	4	222.25	4.5	yes	-
441	19896511	0.608	0	2	402.57	6.9	no	-
442	9969520	0.608	4	6	372.47	4.5	yes	-
443	17828	0.606	2	5	388.5	1.8	yes	13990278
444	19977915	0.606	5	6	470.69	5.9	no	22869502
445	10897138	0.605	2	5	576.81	8.3	no	-
446	161375	0.604	0	4	335.44	N.A.	-	10449622
447	10064504	0.603	1	2	318.49	5.8	no	-
448	10569409	0.602	1	3	343.5	4.7	yes	-
449	10838865	0.602	0	3	492.73	7.7	no	-
450	19896488	0.601	0	2	402.57	7.5	no	-

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
451	20848952	0.599	4	7	402.5	N.A.	-	-
452	10553104	0.596	1	3	490.8	9.5	no	10553065 10814910
453	9819626	0.593	2	2	346.55	5.5	no	-
454	11486584	0.592	2	5	406.56	4.9	yes	-
455	21725453	0.591	1	6	570.78	7.6	no	-
456	11867729	0.588	0	4	348.48	N.A.	-	11867731 11867733 11867735
457	19695788	0.587	2	5	434.61	5.4	no	22882768
458	19871086	0.585	1	3	390.6	7.5	no	-
459	2825513	0.581	1	3	346.5	5.1	no	5110937 11859746 11890335 11890336 11890337 11890338 24105858
460	10988198	0.579	0	2	234.33	4.7	yes	-
461	6439010	0.576	3	6	618.84	8.1	no	-
462	4312809	0.575	0	3	566.9	10.8	no	-
463	20347705	0.569	1	4	439.03	5.7	no	22810353
464	6810452	0.567	2	4	319.44	4.2	yes	9562302
465	20713802	0.565	1	2	360.57	6.2	no	-
466	2748205	0.561	2	2	396.61	7.1	no	2748153 11867903 11867904 11867905 11867906 11868055 11868056 11868057 11868058
467	621344	0.560	0	5	406.56	5.2	no	22295401
468	16400198	0.559	0	3	566.9	10.8	no	18397905
469	4172046	0.557	1	4	376.53	4.8	yes	-
470	11772434	0.552	1	3	348.52	5.7	no	-
471	613633	0.552	1	2	444.73	8.5	no	15542397 22294866
472	18978482	0.551	1	2	314.46	4.2	yes	-
473	11697749	0.550	1	8	433.43	3.1	yes	-
474	21394895	0.544	0	3	548.5	9.2	no	-
475	544496	0.543	0	2	437.1	8.9	no	-
476	22900389	0.543	2	5	432.59	5.1	no	-
477	23509359	0.540	2	2	430.71	7.9	no	-
478	10647308	0.540	4	6	452.63	5.9	no	19977939
479	9953948	0.540	2	8	421.4	N.A.	-	-
480	23425553	0.540	2	2	402.65	7.4	no	-
481	10051541	0.539	1	2	496.78	11.4	no	10097370
482	20637854	0.538	2	6	474.67	3.2	yes	22132232
483	20372630	0.537	0	4	374.51	3.6	yes	-
484	21310310	0.536	2	3	402.61	5.9	no	23496927
485	14779100	0.536	1	3	304.42	2.8	yes	18599499
486	13132652	0.535	1	4	418.61	6.1	no	-
487	21486348	0.534	1	3	486.77	8.9	no	-
488	21581486	0.528	2	4	443.66	6.1	no	-
489	255263	0.526	1	2	288.42	3.8	yes	-
490	633529	0.526	1	3	446.71	9	no	-
491	23557698	0.524	0	2	346.55	7	no	-
492	4615559	0.519	2	7	446.53	2.1	yes	15030576
493	544228	0.519	1	3	288.38	0.6	yes	1757291 9835602 11873454
494	20289393	0.516	1	3	362.55	7.2	no	-
495	10371124	0.514	5	7	586.76	5.9	no	-
496	363779	0.513	2	3	317.47	3.9	yes	494706 22210057
497	21310343	0.512	1	4	471.7	5.3	no	23496922
498	21677701	0.507	0	2	430.71	9.9	no	-
499	370991	0.500	2	6	484.62	N.A.	-	-
500	20372523	0.500	1	2	304.47	5.3	no	-

1: PubChem Compound ID, 2: predicted probability, 3: number of H bond donor, 4: number of H bond acceptor, 5: molecular weight, 6: whether a chemical compound satisfies Lipinski's rule of five (# of H bond donor $\leq$ 5 & # of H bond acceptor $\leq$ 10 & molecular weight $\leq$ 500 & xlogp $\leq$ 5), 7: not available because some values were not available.

Supplementary Table A44 Details of predictions in Fig. III.16C

	[1] ID	[2] prob.	[3] H don.	[4] H accept.	[5] M.W.	xlogp	[6] Lipinski5	compounds of same connectivity
1	11450853	0.991	1	2	346.55	6.2	no	-
2	23599904	0.991	1	2	320.51	6.7	no	-
3	20460242	0.989	1	2	304.47	4.6	yes	-
4	10091357	0.989	1	2	383.36	5.5	no	-
5	11698085	0.989	1	7	449.49	4	yes	-
6	21399446	0.988	2	4	362.5	4.4	yes	-
7	9876052	0.988	2	6	869.31	[7] N.A.	no	-
8	9824174	0.987	1	8	433.43	3.8	yes	11690589 18753307
9	11223329	0.986	1	2	400.64	6.6	no	11749960
10	20092534	0.986	1	2	318.49	5.2	no	-
11	11697749	0.986	1	8	433.43	3.1	yes	-
12	16627134	0.984	1	2	360.57	6.7	no	-
13	131285	0.984	1	2	292.46	5.2	no	21121414
14	4157657	0.983	1	3	315.45	3.7	yes	-
15	21511858	0.983	1	3	336.51	N.A.	-	-
16	9995441	0.983	1	3	308.43	3.8	yes	10267286
17	12002529	0.982	1	4	500.75	9.5	no	-
18	11539442	0.982	1	7	403.4	3.2	yes	-
19	9802277	0.981	1	8	421.39	3.4	yes	18753335
20	11163336	0.980	1	2	330.5	5	no	11198308 11370933 11739019
21	11719005	0.980	1	8	433.43	3.8	yes	-
22	11611466	0.979	1	7	419.4	2.8	yes	-
23	2748117	0.979	2	4	350.49	5.5	no	11867730 11867732 11867734 11867736
24	11532117	0.978	1	8	433.43	3.1	yes	-
25	18753324	0.977	1	7	437.85	3.9	yes	-
26	9826529	0.977	1	7	482.3	4.1	yes	18753320
27	11575793	0.977	2	7	439.86	N.A.	-	18753348
28	11539441	0.976	2	7	439.86	N.A.	-	-
29	20643959	0.975	1	1	304.51	6.5	no	-
30	11539440	0.975	1	7	403.4	3.9	yes	-
31	3008386	0.972	2	5	247.21	1.9	yes	-
32	20547436	0.972	1	2	318.49	5.8	no	22819626
33	11718641	0.972	2	3	416.64	5	no	-
34	9868174	0.971	2	3	446.44	N.A.	-	-
35	22729961	0.971	2	3	318.45	3.1	yes	-
36	6011	0.971	1	2	304.47	4.8	yes	3167 12446741 16627148 22794328 22821231 23615698 23615699
37	11176230	0.970	1	2	372.58	6.7	no	-
38	150854	0.970	1	2	288.42	3.8	yes	10039774 10085744 14324933 21122962
39	623740	0.970	2	3	304.42	1.8	yes	150968 1149869 1778898 1778899 1778901 6546064 11872739 11872740 16395893 22849351
40	11575794	0.969	1	7	403.4	3.2	yes	18753349
41	600744	0.969	0	3	563.65	9.3	no	-
42	21312299	0.968	1	4	310.11	3.5	yes	21312290
43	715670	0.968	1	4	265.66	3.4	yes	-
44	21917762	0.967	1	4	423.56	N.A.	-	-
45	18753323	0.967	1	8	445.44	3.1	yes	-
46	11690220	0.967	1	7	417.43	3.4	yes	-
47	12306723	0.966	1	3	458.72	8.5	no	-
48	23495	0.966	2	3	352.9	2.1	yes	235091 3427265
49	2756	0.966	3	6	252.34	1	yes	-
50	5878	0.966	1	3	306.44	4.2	yes	249098 6429885 6432530 9883009 23616084
51	2375	0.966	2	9	430.37	2.5	yes	56069 441405 16110572
52	10203	0.966	1	2	290.44	4.3	yes	225 5879 5880 11303 247732 441302 736280 908626 1274488 1769120 1769123 2754097 2754130 2825511 5318140 5318590 5701989 6420080 6432531 6432554 6603721 10732277 10780013 11055399 11109055 11243255 11437899 11868646 11868647 11868648 11868649 11868766 11868767 11868768 11868769 11869362 11869363 11869364 11869365 11875890 11890326 11890327 11957461 12306761 12306774 16401171 16758042 16759306 23729586 24131069
53	2914	0.966	0	4	416.94	3.2	yes	9880 3034800 5702030 6603774 6714003 9823237 9853228 16757655 16759143

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
54	21563805	0.966	0	4	508.39	3.6	no	-
55	10632	0.966	2	2	304.47	4.1	yes	229021 521368 1547492 1715083 1806551 1806552 1806553 6429878 6432516 7060892 7092617 7092618 9796494 11869526 11869527 16401508 22211700 23857193
56	10634	0.966	2	2	290.44	3.6	yes	102192 223407 968807 1778952 1778953 9814541 9904098 9922062 9965812 10062826 11872752 11872753 11872754 13215813 16395906 18444027
57	3387	0.966	2	4	336.44	2	yes	6446 6429867 6432701 10088392 20056778
58	4435	0.966	0	3	406.56	5.6	no	229455 11869560 11872989 16394511 16396215 23615856
59	2735	0.966	1	1	384.64	7.5	no	6221 1548921 5280795 5283710 5283711 5283712 5353527 5363362 6393768 6432644 6604201 6604662 6708595 6713938 6992015 6992016 7067439 7067440 7251172 7251174 9821465 10000117 10045875 10340013 10883523 10894379 11014566 11025493 11058152 11463269 12303103 17756775 20849436 21304335 22811020 22862325
60	3397	0.966	1	6	276.21	2.6	yes	10802904
61	3170	0.966	0	3	360.53	5.8	no	224004 10316496 16627147 20056767 23615700
62	5833	0.966	0	4	416.57	3.4	yes	5267 941624 1255733 1425524 1742266 1742267 2825513 5110937 6419955 6432512 6604008 6604182 6710654 10477021 11509929 11859746 11863525 11869418 11869419 11869420 11869578 11869579 11869580 11890335 11890336 11890337 11890338 11957686 16394537 16757735 16760181 18594034 24105858
63	4493	0.966	1	7	317.22	1.8	yes	-
64	4086221	0.966	2	6	412.54	2.5	yes	-
65	15	0.966	1	2	290.44	4.5	yes	10635 11302 147320 667503 968803 1779160 1779163 1779165 2748143 6432518 6951009 9965809 10379482 10780012 10902374 11860392 11867857 11867858 11867859 11867860 11869358 11869359 11869360 11869361 11914201 11920866 12446748 12446752 16111632 16395945 23729584
66	193715	0.966	0	5	400.48	2.7	yes	10001057
67	561925	0.966	1	2	318.49	5.1	no	2754136 11868792 11868793 11868794 11868795 14681481 22215019
68	5881	0.966	1	2	288.42	3	yes	76 134506 719264 908452 1780096 1780097 2817714 3036244 5318142 6432532 6603817 7048561 7092995 9814492 9860744 10085638 10541387 10613370 11889770 12358559 12358563 16401868 16759247 23968375
69	10204	0.966	1	2	288.42	3.6	yes	5408 6013 186663 242356 369323 638016 711503 908606 908607 908608 1312122 2825665 3034651 5701998 6432567 6541850 7048589 10063008 10108190 10469710 11514789 11861311 11862155 11869421 11869422 11869423 11870497 11890585 11890586 11920270 12304539 12304540 12304570 16220011 16758160 19615208 20848917 22876164
70	6128	0.966	0	2	286.41	3	yes	2193 249335 517670 712379 719687 1777653 1777654 1777655 2825552 6432604 6603719 7092858 7092862 10636976 11862468 11869385 11869386 16401731 22211566
71	11633120	0.965	2	7	419.4	2.6	yes	-

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
72	19027438	0.964	3	5	362.46	1.4	yes	-
73	21532454	0.963	1	3	299.24	N.A.	-	-
74	10341837	0.963	1	2	416.34	4.5	yes	-
75	11130951	0.962	2	3	318.45	3	yes	-
76	11274	0.962	2	3	320.47	4.1	yes	5028851
77	18753330	0.962	1	9	446.4	N.A.	-	-
78	10007408	0.962	2	8	538.28	3.4	no	16110566 16110573
79	10542856	0.961	1	3	308.43	3.8	yes	-
80	20301641	0.961	2	2	304.47	4	yes	-
81	3084142	0.960	0	3	416.64	6.8	no	13014314 13014317
82	130408	0.960	2	4	346.46	1.8	yes	21121280
83	17896824	0.959	2	2	402.65	7.3	no	-
84	10005449	0.959	1	7	482.3	4.1	yes	23082061
85	11532211	0.958	1	7	437.85	3.9	yes	-
86	5284279	0.958	1	2	402.65	7.2	no	-
87	10181894	0.957	0	3	442.67	7.5	no	-
88	10141010	0.954	1	4	486.73	9.3	no	-
89	592694	0.953	1	2	238.37	4.8	yes	-
90	11163334	0.952	1	2	330.5	5	no	11186567 11290455 11484329
91	5146437	0.952	2	4	332.43	2.8	yes	-
92	10551934	0.952	1	4	460.69	7	no	-
93	9954749	0.951	2	7	437.46	4.8	yes	10717833
94	9934361	0.951	2	9	469.46	3.5	yes	-
95	20301524	0.951	2	3	318.45	2.3	yes	22797425
96	10005226	0.950	2	3	476.77	8.9	no	10027981 10028102 10367854
97	397933	0.949	3	3	458.72	6.9	no	9825415
98	11336944	0.949	1	2	346.55	5.9	no	-
99	21329128	0.949	2	3	304.42	2.2	yes	-
100	11539439	0.948	2	7	439.86	N.A.	-	-
101	10086403	0.947	1	2	304.47	4.8	yes	10335214 10892064 11833657
102	6428271	0.947	0	3	360.53	6.1	no	6428245 6428254 6428262 20837712 20837718
103	242500	0.947	2	4	346.46	1.1	yes	3957732
104	18753367	0.947	1	8	421.39	3.4	yes	-
105	48619	0.946	0	3	344.49	5.5	no	-
106	5090736	0.946	0	3	346.5	5.5	no	-
107	20669540	0.945	1	3	496.76	7.1	no	-
108	11689938	0.944	1	8	404.39	2.1	yes	-
109	43373	0.942	1	2	330.5	4.9	yes	4068828 9797605 16627138 22798805
110	23134402	0.941	1	8	414.37	3.5	yes	-
111	13132686	0.941	1	4	404.58	5.7	no	-
112	11313131	0.940	1	2	316.48	4.7	yes	11483925
113	11312873	0.939	2	3	308.43	3.2	yes	-
114	3577208	0.939	3	4	445.62	3.2	yes	10527325 16399356
115	9869358	0.938	1	10	471.4	4.2	yes	11155655 18753314
116	631387	0.938	0	3	470.73	8.6	no	-
117	18753361	0.938	1	10	471.4	4.2	yes	-
118	23931165	0.937	1	2	274.4	3.2	yes	-
119	418220	0.936	2	4	574.83	5.9	no	21144389
120	11512907	0.935	1	5	628.88	11.2	no	-
121	21720856	0.935	0	2	360.57	7.8	no	-
122	440678	0.934	1	2	402.65	7.4	no	5118589 5284270 15955969 15955970
123	13084790	0.933	1	2	386.61	5.9	no	-
124	23365287	0.933	4	6	442.64	5.6	no	-
125	23724680	0.932	2	3	430.66	5.5	no	-
126	227058	0.930	1	2	318.49	5.3	no	16091784 20599277
127	23504659	0.929	3	4	432.64	4.4	yes	-
128	3328924	0.928	1	1	303.46	N.A.	-	-
129	20825150	0.927	1	4	418.61	5.7	no	-
130	21494700	0.927	1	2	300.44	3.7	yes	-
131	620925	0.926	1	4	376.53	4.5	yes	10237308 11617759 22295371 22295372
132	125155	0.925	2	3	416.64	5.4	no	21120484
133	10393492	0.925	1	3	584.61	9.3	no	-
134	20586756	0.924	2	9	430.37	N.A.	-	-
135	20301529	0.924	2	3	318.45	2.3	yes	-
136	11661370	0.923	1	8	404.39	2.9	yes	-
137	3044055	0.923	0	4	638.92	12.6	no	-
138	15126713	0.923	1	3	316.43	1.1	yes	19418208
139	97789	0.922	1	2	318.49	5.2	no	-
140	23902336	0.922	2	4	336.47	5.2	no	-
141	22895751	0.921	1	3	390.6	6.7	no	-
142	21485543	0.920	1	2	364.52	5.2	no	-
143	20372464	0.918	2	2	318.49	4.7	yes	-
144	20525532	0.917	3	4	334.45	3	yes	-

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
145	13132648	0.917	2	3	348.52	4.5	yes	-
146	10095698	0.915	0	3	458.72	8.9	no	-
147	541358	0.915	0	2	416.68	9.5	no	-
148	567452	0.913	0	2	430.71	9.7	no	11122885 11144488
149	237186	0.912	1	2	318.49	5.2	no	21117221 21117222
150	19977887	0.911	5	6	430.63	5.2	no	22869504
151	4179980	0.910	1	2	288.42	3.8	yes	-
152	15227104	0.909	1	2	369.34	4.9	yes	20564133
153	21436611	0.909	1	2	324.89	4.6	yes	22797535
154	14283994	0.907	0	3	411.37	5.6	no	14283993 15965727
155	628915	0.905	0	2	316.48	5.8	no	-
156	288383	0.905	1	2	306.48	6.3	no	-
157	23400016	0.904	1	3	446.71	8.5	no	-
158	10335459	0.904	2	3	308.43	4.4	yes	10403027
159	539857	0.903	1	2	238.37	3.3	yes	193199 10466745 14138887 23426590
160	15444	0.902	1	3	308.43	4.4	yes	242489 519277 21139712
161	18655913	0.902	1	2	302.45	4.1	yes	18959518
162	10166889	0.901	2	5	572.86	7.5	no	-
163	618872	0.900	1	2	318.49	4.8	yes	-
164	320698	0.897	4	8	636.91	N.A.	no	-
165	252287	0.897	0	4	350.47	5.2	no	239219 11954113 23234428
166	9861311	0.896	3	4	305.41	2	yes	-
167	9932320	0.894	1	8	428.41	3	yes	11697622 18753328
168	10205718	0.893	1	3	474.67	7.3	no	-
169	5255020	0.893	1	4	464.64	6.8	no	-
170	4065913	0.891	2	3	334.49	5.1	no	11877511
171	11144484	0.891	2	3	430.66	6.8	no	-
172	21437899	0.890	1	4	362.5	6.2	no	-
173	20848953	0.890	2	3	304.42	3.5	yes	21252241 21252246
174	461263	0.889	2	7	574.79	7.9	no	-
175	3542089	0.888	2	5	496.68	8.2	no	-
176	247863	0.887	1	2	304.47	4.6	yes	244128
177	573668	0.887	1	2	288.42	4.2	yes	-
178	620781	0.887	1	3	390.6	6.7	no	2754263 11888201 11888202 11888203 11888204 21679300 21679301
179	21032949	0.886	1	4	380.49	7.3	no	-
180	631371	0.885	0	3	470.73	8.5	no	-
181	13084788	0.885	1	2	372.58	5.3	no	-
182	10838865	0.884	0	3	492.73	7.7	no	-
183	119046	0.884	4	6	499.7	3.7	yes	9935704 13955658
184	23599901	0.883	1	2	320.51	6.8	no	-
185	11795274	0.883	0	4	400.55	6.3	no	-
186	6422127	0.880	2	4	398.58	4.9	yes	-
187	17789823	0.880	2	5	376.49	2.1	yes	-
188	18988333	0.879	3	6	634.88	N.A.	no	-
189	21711974	0.878	1	4	409.54	N.A.	-	-
190	23504636	0.878	3	5	502.73	5.7	no	-
191	18649213	0.876	2	4	346.46	2.5	yes	18947468 21118288
192	11877510	0.876	1	3	333.48	N.A.	-	-
193	11560992	0.875	1	2	428.69	7.4	no	15940327
194	17814850	0.875	2	8	418.42	3.1	yes	-
195	287693	0.873	0	3	360.53	6	no	-
196	6336041	0.873	2	5	507.72	N.A.	no	-
197	351789	0.870	1	3	332.48	5.1	no	-
198	5201680	0.870	1	3	358.51	5.8	no	14863152 20055616
199	11614288	0.870	1	5	628.88	11.2	no	-
200	16731637	0.869	1	2	274.4	3.2	yes	-
201	11825674	0.869	1	2	406.6	5.7	no	-
202	21019537	0.869	3	8	640.85	7.9	no	-
203	23365193	0.867	4	6	442.59	3.9	yes	-
204	20317308	0.867	1	3	342.47	4.6	yes	-
205	273381	0.866	1	2	304.47	5.1	no	-
206	225477	0.865	2	6	628.92	7.7	no	3545386 22793133
207	222791	0.864	2	3	290.4	1.6	yes	3827295
208	9996510	0.861	0	2	326.47	4.2	yes	-
209	11675775	0.860	1	8	404.39	2.9	yes	-
210	16744990	0.859	1	4	364.52	3.5	yes	-
211	17898263	0.858	2	12	552.42	N.A.	no	-
212	22296473	0.857	1	3	432.68	8.1	no	22296474
213	19896511	0.856	0	2	402.57	6.9	no	-
214	11103275	0.854	0	1	515.61	10.7	no	-
215	21771350	0.854	1	5	454.66	6.7	no	-
216	10892705	0.851	0	1	324.5	5.4	no	20713641
217	9578205	0.849	2	4	604.95	9.4	no	-

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
218	11268014	0.849	1	2	346.55	5.8	no	-
219	245623	0.849	0	3	456.7	9.4	no	-
220	20574214	0.847	2	3	310.45	3.9	yes	-
221	20274651	0.847	1	2	469.2	N.A.	-	22826685
222	301438	0.847	1	2	332.52	5.7	no	-
223	23365256	0.847	3	6	456.66	6.1	no	-
224	10393991	0.847	1	3	612.92	12.3	no	-
225	20319857	0.845	0	1	242.78	4.8	yes	-
226	21500111	0.843	1	5	365.46	N.A.	-	-
227	14259153	0.843	2	3	304.42	1.9	yes	23256526
228	10723127	0.841	3	5	670.02	9.8	no	-
229	23496841	0.840	1	2	382.58	6.9	no	-
230	9849601	0.840	1	5	521.69	5.7	no	10324425
231	23504603	0.839	0	3	369.52	N.A.	-	-
232	20325529	0.838	5	2	303.28	N.A.	-	-
233	10320641	0.835	1	8	433.43	3.2	yes	23082062
234	13833963	0.835	2	3	328.45	1.7	yes	18595696
235	19827354	0.833	3	5	362.46	2.6	yes	-
236	22709408	0.832	3	4	362.5	2.7	yes	-
237	630636	0.831	1	4	406.6	5.6	no	22296332
238	9952046	0.830	1	4	384.55	4.6	yes	10200056 21709983
239	11317146	0.829	1	4	460.69	8.2	no	-
240	5039651	0.829	1	3	436.63	7.7	no	5704602 11886946 11886947 11886948 11886949
241	13188693	0.829	1	3	288.38	0.6	yes	-
242	11678106	0.829	3	12	526.48	N.A.	no	-
243	11039070	0.828	2	6	568.78	8.1	no	-
244	21680334	0.827	0	4	488.74	9.7	no	-
245	251593	0.825	2	3	304.42	2.9	yes	4038155
246	23203409	0.825	2	5	435.6	5.1	no	-
247	9980385	0.823	2	5	434.61	5.4	no	-
248	3012008	0.823	5	6	807.54	6.5	no	-
249	10886169	0.823	1	4	553.61	7.4	no	11006178
250	625870	0.822	0	2	290.44	5.5	no	-
251	22159025	0.820	1	2	230.73	4.3	yes	-
252	22159024	0.820	1	2	275.18	4.6	yes	-
253	244915	0.818	0	3	318.45	4.7	yes	4545683
254	397626	0.818	3	12	963.29	12.2	no	500422 22210664
255	11236199	0.818	2	3	446.71	7.7	no	14284349 23427327
256	21711978	0.817	1	5	447.63	N.A.	-	-
257	11432487	0.816	1	1	526.58	9.9	no	11656613
258	10841402	0.816	4	7	608.8	8.3	no	-
259	22082641	0.816	0	4	416.57	3.6	yes	-
260	10456049	0.813	1	3	442.67	7.8	no	-
261	15560268	0.812	2	3	416.64	5.6	no	21669871 23425493 23425495
262	23724589	0.809	1	2	416.68	7.7	no	-
263	10257971	0.807	1	7	597.78	9	no	-
264	20637854	0.806	2	6	474.67	3.2	yes	22132232
265	13917506	0.803	2	5	376.49	4.4	yes	-
266	23509396	0.801	2	2	430.71	6.8	no	-
267	11045841	0.801	0	3	332.48	5.6	no	-
268	623134	0.798	1	2	416.68	7.7	no	22295526 22295527
269	10556204	0.795	0	4	648.98	12.6	no	-
270	9836032	0.795	2	3	302.41	3.3	yes	9922452 9944238 9944239 22726705
271	10554608	0.795	1	5	546.8	9.3	no	-
272	158469	0.795	1	2	330.5	5.4	no	-
273	228493	0.794	1	5	418.57	5	no	3841566
274	3129307	0.794	1	3	572.86	10.5	no	5762860 16396408
275	10589604	0.790	1	2	290.44	4.3	yes	-
276	2807124	0.789	3	5	423.54	5.3	no	3714721 5714227 11889379 11889381 11889383 11889385
277	4655293	0.789	2	4	384.51	3.2	yes	6530688
278	20521976	0.786	3	5	349.46	3.6	yes	-
279	10592486	0.786	1	2	329.52	5.3	no	-
280	11529815	0.785	1	3	318.45	2.6	yes	-
281	10271437	0.783	1	5	385.54	5.3	no	21710003
282	6438783	0.780	0	0	400.7	8.7	no	20839175
283	19609537	0.778	3	6	624.87	N.A.	no	-
284	23720029	0.778	1	4	428.65	N.A.	-	23719675
285	152682	0.776	2	5	372.52	4	yes	-
286	23687034	0.776	1	4	412.54	N.A.	-	-
287	21485685	0.775	2	2	314.46	4.1	yes	-
288	5272689	0.772	2	3	444.69	7.2	no	15488955 15488956
289	23322667	0.769	2	3	374.56	4.8	yes	-
290	19977932	0.767	5	6	456.66	5.5	no	22869503

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
291	10378880	0.767	1	2	279.36	3.2	yes	-
292	63023	0.766	1	2	300.44	3.9	yes	623668 14260703 22295622 23586069
293	11742322	0.764	0	1	410.67	8.3	no	21470113
294	22042398	0.763	2	3	550.85	7.9	no	-
295	222813	0.763	2	3	346.5	3.4	yes	-
296	20545999	0.762	2	3	318.45	2.4	yes	-
297	572677	0.762	1	1	290.48	6.3	no	22215819 22215820
298	11834066	0.761	1	2	316.48	4.5	yes	-
299	126511	0.760	5	9	552.7	5.4	no	10392863
300	13786027	0.757	0	2	300.44	3.6	yes	-
301	9601475	0.757	0	3	443.7	9.8	no	20845308
302	128131	0.756	1	1	396.65	7.4	no	21120944 23425782
303	261357	0.755	1	5	558.81	8.7	no	-
304	21485488	0.753	1	2	302.45	4.3	yes	-
305	345453	0.751	1	2	168.23	3.3	yes	-
306	13982083	0.748	1	2	488.74	8.9	no	18634628 18634629
307	20317310	0.747	1	5	402.52	4.3	yes	-
308	19454	0.747	1	2	302.45	4.1	yes	251636 4619902 23618127
309	10139347	0.747	2	3	458.72	9.5	no	11712271
310	21711954	0.745	1	2	312.45	3.5	yes	23616850
311	19896488	0.744	0	2	402.57	7.5	no	-
312	11802364	0.744	6	10	648.82	5.2	no	-
313	3085068	0.743	2	7	466.59	3.2	yes	21155972
314	19977897	0.742	4	7	448.64	4.4	yes	-
315	604951	0.742	1	1	426.72	9.4	no	10093995 21585583
316	11374078	0.739	1	1	440.74	10	no	-
317	21633176	0.738	0	5	530.78	7.5	no	-
318	2748164	0.737	2	2	290.44	3.1	yes	3970063 11867931 11867932 11867933 11867934
319	10204874	0.736	1	2	460.69	7.2	no	-
320	23365176	0.731	3	6	430.63	5.4	no	-
321	18344556	0.731	2	3	374.56	4.8	yes	18372186
322	19386002	0.730	1	3	464.68	7.5	no	22856060
323	11384476	0.729	0	4	415.61	N.A.	-	20683741 21703528
324	13951	0.727	2	2	302.45	4.4	yes	-
325	243968	0.724	0	3	318.45	4.4	yes	3383374
326	18653140	0.723	0	3	457.04	7.9	no	23358762
327	20344915	0.722	1	5	362.46	2.7	yes	22810486
328	10573794	0.719	2	4	413.6	N.A.	-	-
329	23358761	0.718	1	3	362.55	5.8	no	-
330	10929620	0.718	0	4	542.81	9.4	no	-
331	5321407	0.717	3	3	458.72	6.6	no	12442850 12442851 21136293 21136295
332	21989803	0.716	0	2	445.43	6.5	no	-
333	20317304	0.716	2	4	374.51	3.7	yes	-
334	14055893	0.713	1	3	446.71	9.3	no	16082395
335	21485828	0.711	1	2	330.5	5.1	no	-
336	22898964	0.711	2	9	462.44	2.8	yes	-
337	11537058	0.711	1	4	320.42	3.3	yes	-
338	10633	0.710	1	2	304.47	4.9	yes	4079 18650 244808 2754139 6429828 6429872 6432517 11868804 11868805 11868806 11868807 11871548 11871549 11871550 11871551 20056595
339	21485584	0.710	0	3	420.58	6.2	no	22790394
340	22239603	0.709	1	3	458.72	8.5	no	-
341	22900387	0.709	2	8	835.2	14.1	no	-
342	10530108	0.708	1	4	524.84	9	no	-
343	10326482	0.706	3	6	616.19	7.3	no	21781241
344	10434273	0.700	0	3	458.37	5.2	no	-
345	23379097	0.698	0	2	386.61	6.7	no	-
346	21544019	0.698	1	2	356.54	4.7	yes	-
347	10009174	0.696	2	5	629.66	8.3	no	-
348	19977834	0.696	4	5	406.6	4	yes	-
349	10521467	0.695	2	3	345.52	5.1	no	-
350	23365267	0.694	3	6	442.64	5.5	no	-
351	160526	0.692	1	3	302.41	0.8	yes	5205999 10425017
352	536509	0.690	1	4	376.53	4	yes	-
353	16433479	0.690	2	2	415.01	6.9	no	17376431
354	10316669	0.690	1	4	363.49	4.4	yes	-
355	20564150	0.689	1	3	564.88	N.A.	no	-
356	4577432	0.688	0	3	448.64	7.5	no	5356103
357	21680322	0.687	1	3	458.72	9.7	no	-
358	4626219	0.685	0	1	386.65	8.9	no	-
359	313072	0.684	0	4	432.64	7.6	no	637635 11070078 13200866 21140623 23731163
360	9602878	0.681	2	4	333.47	3.8	yes	20845595 20845596
361	10077305	0.680	0	5	600.57	7.9	no	-
362	5284024	0.678	3	5	404.54	2.8	yes	-
363	9870455	0.678	2	4	496.72	N.A.	-	-

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
364	20372415	0.678	2	3	346.5	3.3	yes	-
365	21019536	0.677	3	8	640.85	7.9	no	-
366	9934521	0.672	2	3	472.74	6.8	no	23504598
367	21355874	0.672	2	5	347.43	N.A.	-	22999989
368	23365189	0.672	4	6	454.65	5.7	no	-
369	20431140	0.670	0	4	346.46	2.9	yes	-
370	10715449	0.670	1	4	390.56	4.9	yes	-
371	21200350	0.667	1	2	298.42	3.6	yes	-
372	21681435	0.664	2	3	265.39	3	yes	21681434 21681436
373	17761457	0.664	2	5	392.58	4.5	yes	-
374	23322668	0.664	1	2	322.87	4.5	yes	-
375	11365422	0.663	1	6	598.85	8.4	no	-
376	23258060	0.662	2	3	444.69	6.1	no	23258061
377	23203431	0.661	3	5	454.6	5.3	no	-
378	21420876	0.659	2	2	346.55	5.6	no	-
379	10448013	0.659	1	3	308.43	4.4	yes	10470542
380	22328355	0.659	1	5	420.58	6.1	no	-
381	22958114	0.657	1	2	330.5	4.6	yes	-
382	11351324	0.656	1	2	446.66	7	no	-
383	412018	0.656	6	8	997.41	10.9	no	-
384	242497	0.655	2	3	318.45	2.1	yes	3246207 4983484 6713733 21157801
385	20521388	0.655	0	5	584.78	10.5	no	-
386	23312556	0.654	4	5	296.22	N.A.	-	-
387	16745166	0.653	1	3	322.46	4.4	yes	16757512
388	601507	0.653	0	3	484.75	8.9	no	-
389	11591257	0.653	2	12	490.38	N.A.	no	-
390	10716928	0.651	2	5	418.61	3.7	yes	-
391	11670426	0.650	1	2	492.78	8.8	no	-
392	18595488	0.649	3	7	791.65	8	no	21564149
393	10196174	0.647	2	2	520.81	N.A.	no	18446459
394	20093901	0.647	2	3	318.45	2	yes	-
395	19064409	0.646	2	4	390.6	6	no	-
396	469170	0.645	3	6	592.81	7.8	no	469168 10031375 10257884
397	162939	0.644	2	6	600.87	6.8	no	18605040 20185899
398	10742644	0.642	1	5	460.63	4.4	yes	-
399	11003996	0.642	1	4	404.58	5	no	11133226 14237013 18641461
400	9966550	0.640	1	2	312.45	3.6	yes	19861705
401	10859758	0.640	2	2	306.48	4.3	yes	-
402	101775	0.640	1	2	400.64	7	no	2752856 2752857 3053399 4995300 11887406 11887407 13285140 21776541
403	10405369	0.639	1	2	346.55	5.6	no	20824888
404	4059540	0.638	1	3	486.77	9.1	no	-
405	15170637	0.638	1	2	304.47	4.3	yes	19832934
406	11732748	0.635	2	2	430.71	8.4	no	21770608
407	20372368	0.635	1	2	288.42	4.5	yes	-
408	11524519	0.635	1	3	416.64	5.8	no	-
409	20355644	0.634	2	3	318.45	3	yes	-
410	21033091	0.634	3	7	598.77	6.7	no	-
411	9929619	0.632	2	5	377.47	2.7	yes	-
412	11969953	0.631	2	5	507.72	N.A.	no	-
413	11250247	0.630	4	10	586.68	8.2	no	-
414	10504787	0.630	3	4	476.73	6.5	no	21582606
415	565634	0.627	0	3	484.75	8.8	no	22215285 22215286
416	623288	0.627	0	3	318.45	4.8	yes	22295563
417	16433480	0.627	2	3	410.59	6.2	no	17376428
418	19603149	0.626	1	4	391.56	4	yes	19603085
419	19603072	0.626	1	4	300.1	3.6	yes	-
420	20358277	0.625	2	3	295.46	4.8	yes	-
421	21410876	0.625	1	2	276.41	4.6	yes	22795658
422	30196	0.622	2	3	334.49	4.5	yes	4305372 23619649
423	20301542	0.621	2	3	332.48	2.8	yes	-
424	127362	0.619	6	6	591.91	7.4	no	-
425	23365197	0.618	4	6	442.64	5.3	no	-
426	20657237	0.617	2	7	689.99	5.7	no	-
427	17885975	0.616	1	3	495.95	8.4	no	-
428	5375219	0.616	2	2	387.39	5	no	23425350 23427916
429	22868528	0.616	1	3	356.5	4.8	yes	23358798
430	9928114	0.615	1	2	344.53	5.6	no	11279391 11450790 22053369
431	20549144	0.615	0	4	437.59	N.A.	-	-
432	18655330	0.614	1	5	452.34	N.A.	-	19418177
433	21709398	0.614	2	3	318.45	2.6	yes	-
434	19418183	0.614	1	5	407.89	N.A.	-	-
435	10852075	0.612	1	2	304.47	4.7	yes	11185797 11278126 23374746
436	18633716	0.612	0	2	432.7	9.5	no	20029438

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
437	449706	0.611	1	3	307.43	4.6	yes	451177 9972568 10448011
438	22817977	0.608	1	3	564.88	N.A.	no	-
439	14064776	0.607	2	10	450.36	2.7	yes	-
440	10315101	0.607	3	4	338.48	3.5	yes	-
441	9916384	0.606	3	5	584.85	8.7	no	-
442	16752677	0.606	2	2	386.61	5.6	no	-
443	6479892	0.606	2	3	444.69	7.3	no	10939190 12996667 13322136
444	134100	0.606	1	2	416.34	5.3	no	-
445	9904	0.606	1	2	274.4	3.2	yes	220503 2754096 6451993 7048722 7087517 10016199 10061922 11865439 11865440 11865441 11865442 11875291 11875292 11888113 11888114 11888115 12358925 16399574 16760199
446	19776159	0.604	0	3	477.76	N.A.	-	-
447	10704282	0.602	1	2	222.32	2.5	yes	-
448	2748151	0.600	1	2	388.54	5.7	no	4272506 11886938 11886939 11886940 11886941
449	21626811	0.599	0	1	400.68	9.5	no	21626813
450	20808114	0.598	0	3	372.52	7.7	no	22384451
451	21327829	0.598	1	3	446.71	8.5	no	-
452	171456	0.597	2	4	388.54	4.1	yes	21124976 21487672
453	19609992	0.595	2	3	304.42	2.6	yes	-
454	13917206	0.594	1	2	332.52	5.1	no	18638064
455	23365264	0.592	4	6	452.63	5.9	no	-
456	258526	0.590	1	2	399.65	6.5	no	3851495 10386236
457	19933655	0.589	0	2	263.4	N.A.	-	-
458	622810	0.584	1	2	414.66	7.3	no	22295493 22295494
459	5316686	0.583	2	4	388.54	4.2	yes	-
460	65543	0.581	2	3	304.42	2	yes	222794 12073407 16219473
461	13084792	0.581	1	2	358.56	6	no	-
462	10572849	0.580	1	4	396.52	5.6	no	-
463	18354302	0.579	9	12	1682.42	N.A.	no	-
464	10118846	0.579	1	5	520.7	6.6	no	-
465	630375	0.578	2	2	332.52	4.4	yes	22296304 22296305
466	10919070	0.578	0	1	597	N.A.	no	-
467	11248818	0.578	1	2	488.86	N.A.	-	11477343
468	20657173	0.577	0	5	570.84	8.9	no	-
469	22044591	0.576	1	1	340.54	7	no	-
470	19977676	0.575	3	7	648.87	N.A.	no	-
471	440368	0.575	2	3	332.48	3.3	yes	107701 625529 5459859 9797681 16219879
472	20779720	0.574	2	3	346.5	2.8	yes	22491655
473	463159	0.573	1	5	444.58	3.6	yes	10003591
474	10504916	0.570	2	2	480.76	10.4	no	10672390 21582717 21582718
475	17978295	0.569	2	6	372.42	3.1	yes	-
476	11867729	0.564	0	4	348.48	N.A.	-	11867731 11867733 11867735
477	20463769	0.564	4	4	350.49	4	yes	22788552
478	22090280	0.563	1	5	532.79	8.6	no	-
479	18464409	0.563	1	4	429.61	N.A.	-	-
480	4775247	0.563	2	4	412.61	5.3	no	-
481	2748093	0.562	1	3	376.57	7	no	4394049 11867630 11867631 11867632 11867633
482	541153	0.560	0	5	484.68	10.9	no	11306270 13409200
483	250804	0.557	0	3	358.51	5.6	no	250803 3436837
484	19977920	0.556	4	6	444.65	5.4	no	22869500
485	19896496	0.554	0	2	458.67	8.6	no	22844093
486	497951	0.553	0	4	362.5	5.3	no	11024869
487	19933656	0.553	1	2	264.4	5	no	-
488	21485576	0.551	1	2	344.53	5.7	no	-
489	20301671	0.550	1	2	302.45	3.6	yes	22803244
490	19064366	0.548	2	4	404.63	6.5	no	-
491	19606895	0.546	2	5	363.45	2.1	yes	-
492	17874621	0.545	1	2	312.45	4.6	yes	-
493	11779579	0.543	1	2	300.44	3.7	yes	-
494	23253348	0.542	0	3	522.8	11.4	no	-
495	10254233	0.541	2	4	472.7	7.7	no	10600528
496	11668856	0.541	1	7	417.43	3.5	yes	-
497	10651262	0.538	2	5	602.84	8.1	no	-
498	247996	0.537	2	4	346.46	1.3	yes	245233 1774727 1774729 1774731 3778676 11876059 11876060 11876061 16401437 18397935
499	22808563	0.537	2	3	328.45	2.3	yes	23331588
500	20301571	0.537	2	3	332.48	2.8	yes	-
501	18959856	0.536	0	3	548.84	12.5	no	22851215
502	21485609	0.535	0	3	406.56	5.8	no	-
503	21355875	0.535	3	5	348.43	3.3	yes	22999990
504	18391386	0.534	0	1	288.47	5.9	no	21219361
505	20372628	0.534	0	3	460.76	N.A.	-	22801498
506	10072691	0.525	1	3	454.68	6.9	no	-
507	20325400	0.525	3	5	362.46	2.3	yes	22791787
508	20029573	0.521	0	1	565.74	12.5	no	-
509	580275	0.520	2	5	418.57	3.1	yes	22216314 22216315

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
510	10642746	0.519	1	3	367.52	5.1	no	-
511	22404988	0.519	2	3	447.05	6.8	no	-
512	9847174	0.519	2	9	462.44	4	yes	-
513	10884228	0.518	0	4	416.59	6.6	no	-
514	21684445	0.518	3	6	678.77	8.9	no	-
515	616684	0.515	2	3	306.44	2.6	yes	11551286 15929496 22295053 22295054
516	248649	0.513	2	3	304.42	3.5	yes	3994025
517	11761566	0.510	3	3	418.65	5.4	no	-
518	10160599	0.509	0	3	442.67	7.5	no	-
519	3546502	0.509	1	2	506.8	10.7	no	-
520	22898963	0.507	2	7	430.44	4	yes	-
521	11530133	0.507	2	4	336.47	3.3	yes	-
522	301437	0.500	1	2	318.49	5.3	no	-
523	3728758	0.500	1	4	452.63	7.1	no	-
524	10018547	0.500	2	2	314.46	4	yes	-
525	10074154	0.500	2	6	486.68	7.6	no	-
526	20564132	0.500	1	3	322.46	4.8	yes	-
527	21419224	0.500	4	7	674.91	N.A.	no	-

1: PubChem Compound ID, 2: predicted probability, 3: number of H bond donor, 4: number of H bond acceptor, 5: molecular weight, 6: whether a chemical compound satisfies Lipinski's rule of five (# of H bond donor $\leq$ 5 & # of H bond acceptor $\leq$ 10 & molecular weight $\leq$ 500 & xlogp $\leq$ 5), 7: not available because some values were not available.

Supplementary Table A45 Details of predictions in Fig. III.16D

	1 ID	2 prob.	3 H don.	4 H accept.	5 M.W.	xlogp	6 Lipinski5	compounds of same connectivity
1	11719005	0.987	1	8	433.43	3.8	yes	-
2	11611466	0.986	1	7	419.4	2.8	yes	-
3	11532117	0.985	1	8	433.43	3.1	yes	-
4	11539737	0.985	1	7	417.43	3.4	yes	-
5	11698085	0.984	1	7	449.49	4	yes	-
6	11539441	0.984	2	7	439.86	7 N.A.	-	-
7	11539442	0.983	1	7	403.4	3.2	yes	-
8	9824174	0.983	1	8	433.43	3.8	yes	11690589 18753307
9	11697749	0.983	1	8	433.43	3.1	yes	-
10	10320641	0.982	1	8	433.43	3.2	yes	23082062
11	11532211	0.982	1	7	437.85	3.9	yes	-
12	17814850	0.982	2	8	418.42	3.1	yes	-
13	11539439	0.981	2	7	439.86	N.A.	-	-
14	11633120	0.981	2	7	419.4	2.6	yes	-
15	11539440	0.981	1	7	403.4	3.9	yes	-
16	18753367	0.981	1	8	421.39	3.4	yes	-
17	11668856	0.980	1	7	417.43	3.5	yes	-
18	11690220	0.980	1	7	417.43	3.4	yes	-
19	11575794	0.980	1	7	403.4	3.2	yes	18753349
20	18753323	0.980	1	8	445.44	3.1	yes	-
21	10005449	0.979	1	7	482.3	4.1	yes	23082061
22	11662271	0.979	1	8	445.44	3.1	yes	-
23	18753361	0.976	1	10	471.4	4.2	yes	-
24	11575793	0.974	2	7	439.86	N.A.	-	18753348
25	11539736	0.973	2	7	453.89	N.A.	-	-
26	18753321	0.972	2	9	447.41	0.7	yes	-
27	9826529	0.969	1	7	482.3	4.1	yes	18753320
28	18753324	0.969	1	7	437.85	3.9	yes	-
29	18753330	0.966	1	9	446.4	N.A.	-	-
30	9802277	0.965	1	8	421.39	3.4	yes	18753335
31	11575288	0.964	1	7	417.43	3.5	yes	-
32	2756	0.963	3	6	252.34	1	yes	-
33	3397	0.963	1	6	276.21	2.6	yes	10802904
34	2748117	0.963	2	4	350.49	5.5	no	11867730 11867732 11867734 11867736
35	4493	0.963	1	7	317.22	1.8	yes	-

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
36	2735	0.963	1	1	384.64	7.5	no	6221 1548921 5280795 5283710 5283711 5283712 5353527 5363362 6393768 6432644 6604201 6604662 6708595 6713938 6992015 6992016 7067439 7067440 7251172 7251174 9821465 10000117 10045875 10340013 10883523 10894379 11014566 11025493 11058152 11463269 12303103 17756775 20849436 21304335 22811020 22862325
37	2914	0.963	0	4	416.94	3.2	yes	9880 3034800 5702030 6603774 6714003 9823237 9853228 16757655 16759143
38	21563805	0.963	0	4	508.39	3.6	no	-
39	5833	0.963	0	4	416.57	3.4	yes	2375 5267 56069 441405 941624 1255733 1425524 1742266 1742267 6419955 6432512 6604008 6604182 6710654 10477021 11509929 11863525 11869418 11869419 11869420 11869578 11869579 11869580 11957686 16110572 16394537 16757735 16760181 18594034
40	11675775	0.962	1	8	404.39	2.9	yes	-
41	162324	0.961	0	4	404.56	2	yes	-
42	22082641	0.960	0	4	416.57	3.6	yes	-
43	193715	0.958	0	5	400.48	2.7	yes	10001057
44	9868914	0.956	2	9	461.44	2.9	yes	-
45	11554146	0.954	2	9	447.41	0.7	yes	-
46	9869358	0.948	1	10	471.4	4.2	yes	11155655 18753314
47	9910546	0.947	2	7	430.44	5.2	no	-
48	11689938	0.946	1	8	404.39	2.1	yes	-
49	11166353	0.945	1	7	432.44	4.4	yes	-
50	4508399	0.942	1	1	364.56	7.1	no	16433532 16831962
51	11518267	0.939	1	9	461.44	3.2	yes	-
52	9868236	0.939	1	8	447.45	4	yes	18753366
53	11867729	0.938	0	4	348.48	N.A.	-	11867731 11867733 11867735
54	9932320	0.937	1	8	428.41	3	yes	11697622 18753328
55	22084311	0.935	0	4	418.59	2.3	yes	-
56	9933226	0.932	2	3	445.68	5.6	no	-
57	9934361	0.926	2	9	469.46	3.5	yes	-
58	11661370	0.926	1	8	404.39	2.9	yes	-
59	10671749	0.913	2	5	462.66	6.4	no	-
60	17769708	0.913	1	2	374.52	4.7	yes	-
61	10007408	0.911	2	8	538.28	3.4	no	16110566 16110573
62	22898971	0.911	2	8	414.38	N.A.	-	-
63	9931281	0.903	1	6	408.44	3.9	yes	18753351
64	16110571	0.902	2	7	479.27	4.6	yes	-
65	21656519	0.902	1	3	416.55	5.3	no	-
66	9955874	0.900	2	3	459.7	6.1	no	-
67	592694	0.884	1	2	238.37	4.8	yes	-
68	18766981	0.877	0	3	402.61	6.9	no	20606261
69	20374764	0.876	1	3	420.62	7.3	no	-
70	240769	0.873	1	3	322.46	3.8	yes	4411881
71	161375	0.871	0	4	335.44	N.A.	-	10449622
72	597417	0.870	0	2	252.39	5.1	no	-
73	22898952	0.864	2	7	437.46	N.A.	-	-
74	23572073	0.860	1	1	392.62	8.1	no	-
75	23535433	0.859	0	3	400.59	6.5	no	-
76	11744096	0.853	1	2	442.72	8.4	no	-
77	23239871	0.850	0	3	472.74	8.8	no	-
78	623809	0.848	0	3	362.55	5.6	no	22295635
79	10390972	0.844	1	1	492.82	11.4	no	10458382
80	17566945	0.843	1	3	341.83	2	yes	-
81	9836509	0.842	1	2	314.46	6.8	no	20569562
82	22042365	0.840	2	3	564.88	8.4	no	-
83	9891033	0.840	2	3	473.73	6.4	no	-
84	16627194	0.840	0	4	430.6	3.6	yes	18477668
85	10292678	0.838	1	5	372.46	6.4	no	17978254
86	441676	0.826	1	2	442.72	7.5	no	466377 4220788 5316940 10411112 18530313
87	17978321	0.826	1	7	422.47	7.1	no	-
88	18605221	0.821	0	2	459.1	9.8	no	-
89	18605220	0.821	0	2	503.55	10.1	no	-
90	18605223	0.819	0	3	442.65	9.6	no	-
91	10908100	0.819	0	3	594.87	12.2	no	-
92	22042398	0.817	2	3	550.85	7.9	no	-
93	565634	0.817	0	3	484.75	8.8	no	22215285 22215286
94	21940268	0.816	2	3	468.69	5	no	-
95	601507	0.815	0	3	484.75	8.9	no	-

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
96	370496	0.813	0	3	388.58	6.5	no	495075 2754284 10949126 11869019 11869020 11869021 11869022 17369966 18869233 22210112 22821830
97	20538550	0.811	1	4	390.54	2.2	yes	22820190
98	600744	0.802	0	3	563.65	9.3	no	-
99	20454298	0.802	1	4	390.54	2.1	yes	-
100	21123131	0.802	0	5	369.5	N.A.	-	21123132 23112953
101	17445779	0.800	2	3	343.85	N.A.	-	-
102	3593323	0.799	1	2	406.6	6.8	no	5869956
103	22898962	0.786	2	8	421.4	N.A.	-	-
104	21263876	0.786	1	3	512.72	8.6	no	-
105	21725580	0.784	0	3	416.64	7	no	21725577
106	16752677	0.782	2	2	386.61	5.6	no	-
107	9911467	0.776	1	8	447.45	4.1	yes	18753341
108	10178233	0.768	1	5	386.49	6.7	no	-
109	11125303	0.764	1	2	685.03	13.6	no	21672038
110	11774947	0.761	1	6	436.86	5.1	no	-
111	11386097	0.761	1	6	481.31	5.2	no	-
112	21940264	0.754	2	3	482.72	5.1	no	-
113	21815135	0.752	1	2	418.61	7.3	no	-
114	5316248	0.750	1	1	442.76	10.1	no	-
115	9806757	0.745	2	3	521.77	7.3	no	-
116	10180572	0.737	1	7	422.47	7.3	no	17978313
117	10160192	0.736	1	7	436.5	7.6	no	-
118	21800710	0.732	1	2	456.74	8.9	no	-
119	23400079	0.730	0	3	402.61	6.8	no	-
120	4250579	0.729	1	1	476.78	10.8	no	-
121	9823440	0.729	1	7	420.4	4.6	yes	11154359 18753350
122	13917206	0.727	1	2	332.52	5.1	no	18638064
123	14064775	0.724	2	8	418.36	3.9	yes	-
124	13039175	0.722	1	2	196.29	4.5	yes	-
125	4456132	0.718	1	3	483.52	7.6	no	-
126	10097724	0.718	1	2	506.8	10.9	no	10413885
127	23524563	0.716	0	4	608.9	13	no	-
128	634100	0.710	1	2	402.65	7.5	no	22296765 22296766
129	10838800	0.705	2	5	490.72	7.3	no	-
130	10477247	0.703	1	2	420.63	7.1	no	11742895
131	244385	0.702	1	5	344.41	4.5	yes	-
132	21520934	0.699	2	3	364.54	2.4	yes	22791793
133	4923947	0.698	3	2	355.86	2.5	yes	-
134	9804987	0.698	3	4	475.7	4.9	yes	-
135	10202445	0.695	1	7	422.47	7.3	no	17978256
136	11661683	0.690	2	8	418.42	2.7	yes	-
137	563499	0.690	1	2	506.8	10.5	no	-
138	23386202	0.688	0	1	398.66	8	no	-
139	1476400	0.688	2	7	421.82	3.3	yes	1476401 3697836
140	9889057	0.683	2	4	432.64	5.8	no	-
141	631387	0.682	0	3	470.73	8.6	no	-
142	9986205	0.682	0	4	610.93	N.A.	no	15978749
143	10051829	0.680	1	2	503.8	10.9	no	10436149 21800673
144	248277	0.679	1	2	362.93	5.6	no	5105189
145	20768371	0.670	0	4	608.9	12.9	no	-
146	9547324	0.669	1	4	450.59	4.9	yes	-
147	10052010	0.667	2	2	508.82	10.4	no	10391562 21800666
148	16657483	0.666	0	5	460.63	3.6	yes	23310560
149	11761768	0.660	3	4	428.6	4.9	yes	-
150	631371	0.657	0	3	470.73	8.5	no	-
151	22898963	0.654	2	7	430.44	4	yes	-
152	11577143	0.654	1	3	513.73	8.3	no	-
153	10367106	0.653	2	5	462.66	7.3	no	-
154	11678106	0.653	3	12	526.48	N.A.	no	-
155	10528599	0.651	2	5	476.69	6.9	no	-
156	11732662	0.646	2	3	426.08	N.A.	-	-
157	6438783	0.642	0	0	400.7	8.7	no	20839175
158	10842404	0.638	3	9	696.95	8.4	no	-
159	10295792	0.629	1	7	422.47	7.4	no	-
160	152305	0.625	1	5	370.5	3.6	yes	18665256 21123133 21848561
161	634454	0.625	1	4	488.74	9.4	no	-
162	19994468	0.623	0	3	548.84	11.1	no	-
163	10549446	0.622	1	2	408.62	6.7	no	10597607 21710040
164	12888731	0.621	1	4	517.53	8.4	no	-
165	9804916	0.617	2	3	473.73	6.5	no	-
166	9969519	0.617	1	5	372.46	6.5	no	-
167	10602001	0.616	1	5	518.77	8	no	-
168	16433527	0.616	1	1	378.59	7.3	no	16831961
169	19362805	0.615	1	5	511.71	N.A.	no	22885103
170	9876052	0.614	2	6	869.31	N.A.	no	-
171	625783	0.613	0	3	376.57	6	no	22295852 22295853
172	17965751	0.612	3	9	447.82	2.9	yes	-

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
173	11476379	0.604	1	2	446.66	6.7	no	11487665
174	2825667	0.603	1	2	392.57	6.7	no	3277179 5716615 5780473 11861313 11890587 11890588 11890589 11890590 11897426
175	23274892	0.598	3	4	265.33	0.3	yes	-
176	10029481	0.596	1	3	514.77	11.5	no	10097997
177	11166513	0.594	1	8	438.39	4.8	yes	-
178	11178122	0.593	1	8	438.39	4.8	yes	-
179	11092132	0.592	0	3	486.77	9.2	no	22296790
180	9547284	0.587	2	5	440.58	6.5	no	-
181	17978280	0.586	1	5	358.44	5.9	no	-
182	23293536	0.578	2	2	414.66	6.8	no	-
183	16681401	0.577	0	2	454.73	8.1	no	-
184	301438	0.576	1	2	332.52	5.7	no	-
185	273381	0.576	1	2	304.47	5.1	no	-
186	10051542	0.574	1	2	496.78	11.4	no	10097369 21800658
187	4437573	0.567	1	2	288.42	3.4	yes	-
188	10365647	0.566	1	2	434.65	7.4	no	-
189	10346339	0.565	1	1	513.24	11.8	no	10369174 21800668
190	10054383	0.565	1	1	604.69	12.3	no	10347708 10415352 11758350 21800692 21800693
191	21399446	0.564	2	4	362.5	4.4	yes	-
192	250207	0.564	1	2	288.42	3.4	yes	-
193	9890426	0.561	1	3	459.7	5.8	no	-
194	9935196	0.557	1	3	487.76	6.7	no	-
195	9961712	0.555	7	9	700.86	N.A.	no	-
196	17071497	0.549	1	5	392.2	5.5	no	-
197	1476447	0.543	2	10	455.37	3.6	yes	1476448 3534970
198	20222769	0.542	1	3	394.59	2.5	yes	-
199	11825076	0.540	1	2	380.56	6.4	no	-
200	95620	0.538	2	3	318.45	4.6	yes	229799 2754099 3034498 5355012 5985098 11449974 11886877 11886878 11886879 11897570 11921419 11921420 21150948 22807432
201	21240247	0.537	0	2	621.73	11.9	no	-
202	9957692	0.531	1	3	501.78	7.1	no	-
203	838171	0.530	2	4	237.25	1.9	yes	-
204	3550468	0.520	1	4	282.31	3	yes	-
205	22042302	0.519	3	4	566.85	6.7	no	-
206	4085195	0.518	1	1	443.46	7.9	no	6521570 19590386
207	246874	0.517	1	2	342.51	5.6	no	-
208	5255288	0.517	1	2	456.74	8.4	no	-
209	11093669	0.517	1	7	685.62	N.A.	no	-
210	4923909	0.511	3	4	310.33	1.5	yes	-
211	15086131	0.510	2	4	572.82	9.6	no	19879155
212	10740933	0.500	1	2	422.64	7.1	no	-
213	16433458	0.500	1	2	382.55	7.2	no	-

1: PubChem Compound ID, 2: predicted probability, 3: number of H bond donor, 4: number of H bond acceptor, 5: molecular weight, 6: whether a chemical compound satisfies Lipinski's rule of five (# of H bond donor $\leq$ 5 & # of H bond acceptor $\leq$ 10 & molecular weight $\leq$ 500 & xlogp $\leq$ 5), 7: not available because some values were not available.

Supplementary Table A46 Details of predictions in Fig. III.19C

	1 ID	2 prob.	3 H don.	4 H accept.	5 M.W.	xlogp	6 Lipinski5	compounds of same connectivity
1	17814850	0.995	2	8	418.42	3.1	yes	-
2	11539737	0.994	1	7	417.43	3.4	yes	-
3	11719005	0.994	1	8	433.43	3.8	yes	-
4	10320641	0.993	1	8	433.43	3.2	yes	23082062
5	11698085	0.993	1	7	449.49	4	yes	-
6	11662271	0.993	1	8	445.44	3.1	yes	-
7	11690220	0.993	1	7	417.43	3.4	yes	-
8	11611466	0.992	1	7	419.4	2.8	yes	-
9	18753323	0.992	1	8	445.44	3.1	yes	-
10	9824174	0.992	1	8	433.43	3.8	yes	11690589 18753307
11	11668856	0.991	1	7	417.43	3.5	yes	-
12	11697749	0.990	1	8	433.43	3.1	yes	-
13	11532117	0.990	1	8	433.43	3.1	yes	-
14	11539441	0.989	2	7	439.86	7 N.A.	-	-
15	11539736	0.988	2	7	453.89	N.A.	-	-
16	18753361	0.988	1	10	471.4	4.2	yes	-
17	11633120	0.988	2	7	419.4	2.6	yes	-
18	18753321	0.987	2	9	447.41	0.7	yes	-
19	11532211	0.987	1	7	437.85	3.9	yes	-
20	11575794	0.986	1	7	403.4	3.2	yes	18753349

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
21	18753367	0.985	1	8	421.39	3.4	yes	-
22	11539442	0.983	1	7	403.4	3.2	yes	-
23	11554146	0.983	2	9	447.41	0.7	yes	-
24	11539439	0.983	2	7	439.86	N.A.	-	-
25	18753330	0.982	1	9	446.4	N.A.	-	-
26	11539440	0.979	1	7	403.4	3.9	yes	-
27	11575793	0.979	2	7	439.86	N.A.	-	18753348
28	11675775	0.979	1	8	404.39	2.9	yes	-
29	10005449	0.978	1	7	482.3	4.1	yes	23082061
30	9869358	0.977	1	10	471.4	4.2	yes	11155655 18753314
31	10385990	0.973	1	5	394.86	2.7	yes	23082018
32	9826529	0.972	1	7	482.3	4.1	yes	18753320
33	18753324	0.972	1	7	437.85	3.9	yes	-
34	21061365	0.971	3	9	447.48	2.2	yes	-
35	9802277	0.968	1	8	421.39	3.4	yes	18753335
36	11689938	0.968	1	8	404.39	2.1	yes	-
37	2756	0.967	3	6	252.34	1	yes	-
38	1476447	0.967	2	10	455.37	3.6	yes	1476448 3534970
39	4493	0.967	1	7	317.22	1.8	yes	-
40	3397	0.967	1	6	276.21	2.6	yes	10802904
41	2375	0.967	2	9	430.37	2.5	yes	56069 441405 16110572
42	17566945	0.967	1	3	341.83	2	yes	-
43	11661683	0.967	2	8	418.42	2.7	yes	-
44	1847754	0.965	2	3	422.33	4.6	yes	-
45	22777708	0.965	2	3	422.33	4.6	yes	-
46	2236640	0.965	3	3	440.34	3.7	yes	-
47	9932320	0.963	1	8	428.41	3	yes	11697622 18753328
48	10118209	0.962	3	8	507.6	3.4	no	-
49	1332900	0.960	3	3	450.35	3.7	yes	-
50	1312039	0.960	3	3	405.9	3.5	yes	-
51	18860124	0.960	3	6	431.51	2	yes	-
52	18724729	0.960	2	5	624.37	6.9	no	-
53	11166353	0.958	1	7	432.44	4.4	yes	-
54	1299225	0.957	2	3	391.26	3.1	yes	-
55	888212	0.957	2	3	346.81	2.9	yes	-
56	17100503	0.957	2	4	324.75	3.6	yes	-
57	22777709	0.957	2	3	422.33	4.6	yes	-
58	11214475	0.954	2	6	515.43	3.4	no	-
59	22081421	0.953	3	6	252.34	N.A.	-	-
60	18679679	0.951	3	9	314.34	N.A.	-	-
61	11661370	0.951	1	8	404.39	2.9	yes	-
62	4136288	0.949	3	5	368.84	3.2	yes	-
63	9868914	0.949	2	9	461.44	2.9	yes	-
64	17187893	0.949	2	6	372.33	2.2	yes	-
65	3425136	0.946	2	5	273.29	1.8	yes	-
66	9868236	0.946	1	8	447.45	4	yes	18753366
67	150172	0.946	3	6	276.36	0.8	yes	-
68	4925246	0.945	3	4	355.39	2.7	yes	-
69	129817	0.942	3	5	227.33	0.8	yes	-
70	21313982	0.940	3	6	242.34	0.7	yes	-
71	22287480	0.940	3	3	336.84	2.5	yes	-
72	12923392	0.939	3	7	272.33	1.8	yes	-
73	4925243	0.939	3	4	424.28	3.7	yes	-
74	2221944	0.939	3	4	468.73	3.9	yes	4925333
75	4925287	0.939	3	4	513.18	4.1	no	-
76	24644210	0.938	2	4	437.55	3.2	yes	-
77	1381902	0.937	3	5	475.54	3.8	yes	-
78	11270938	0.933	2	8	453.2	4.2	yes	-
79	9914664	0.928	4	7	521.31	3	no	-
80	16798843	0.927	2	7	420.38	5	no	-
81	11514791	0.926	0	3	290.05	2.6	yes	-
82	23342730	0.925	3	6	295.36	2.3	yes	-
83	9930574	0.925	3	7	395.4	2.8	yes	10715687 11794987
84	20550120	0.924	3	6	257.31	0.9	yes	-
85	23274617	0.923	3	5	253.32	0.2	yes	-
86	24656514	0.923	2	5	455.37	3.9	yes	-
87	13559291	0.922	1	3	429.59	4.5	yes	-
88	4925483	0.922	4	7	474.45	2.4	yes	-
89	18860108	0.921	2	7	442.42	3.9	yes	-
90	21061385	0.921	3	9	461.51	2.7	yes	-
91	21060767	0.920	3	9	496.56	2.1	yes	-
92	3566788	0.919	2	3	406.33	4.7	yes	-
93	11575288	0.918	1	7	417.43	3.5	yes	-

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
94	21060655	0.915	2	9	482.54	4.2	yes	-
95	22156347	0.914	5	5	312.26	N.A.	-	-
96	18571388	0.914	2	3	296.32	2.3	yes	-
97	24517368	0.914	1	5	424.51	4.4	yes	-
98	18955802	0.912	2	6	276.36	1.2	yes	-
99	1287186	0.910	3	3	484.79	3.9	yes	3335307
100	873683	0.905	1	3	349.21	4.2	yes	-
101	131743	0.904	4	5	270.35	0	yes	-
102	19610619	0.904	3	6	420.89	4.5	yes	-
103	9954749	0.903	2	7	437.46	4.8	yes	10717833
104	10377582	0.902	2	6	253.32	0.9	yes	-
105	17110416	0.898	1	2	468.21	5.1	no	-
106	3376316	0.896	1	4	341.38	3.2	yes	-
107	2224777	0.895	4	4	414.48	2.5	yes	-
108	10181959	0.893	1	3	443.62	5.1	no	13298284
109	20458639	0.893	2	6	290.39	1.8	yes	-
110	5743626	0.892	5	6	325.26	N.A.	-	-
111	21313967	0.891	4	4	331.33	N.A.	-	-
112	20348566	0.891	2	4	514.23	N.A.	no	-
113	11510615	0.890	2	11	447.29	3	no	-
114	20373720	0.888	3	5	241.36	1	yes	-
115	20464601	0.887	2	9	320.34	2.4	yes	-
116	4952146	0.887	3	4	383.44	3.6	yes	-
117	1080671	0.884	2	5	417.5	2.2	yes	-
118	11016786	0.883	0	11	521.46	N.A.	no	-
119	9869286	0.883	2	9	469.46	3.5	yes	-
120	19748631	0.882	5	7	306.82	N.A.	-	-
121	5219671	0.881	1	5	339.35	3.1	yes	-
122	3334438	0.879	3	3	450.35	3.3	yes	3334441
123	3334437	0.879	3	3	405.9	3.1	yes	-
124	10095738	0.878	1	6	459.54	5.3	no	17888285
125	8063774	0.878	1	3	384.49	3	yes	8063775
126	17985929	0.877	2	3	429.59	4.3	yes	-
127	1371134	0.876	1	3	349.21	4.2	yes	-
128	3624727	0.875	1	5	339.35	3.1	yes	-
129	24848270	0.871	3	5	277.37	N.A.	-	-
130	9849212	0.869	1	4	510.2	5.3	no	18753319
131	1296999	0.869	3	3	450.35	3.3	yes	-
132	1297997	0.869	3	3	405.9	3.1	yes	-
133	17204774	0.868	3	3	492.43	4.7	yes	-
134	12918615	0.864	3	3	255.36	0.4	yes	-
135	21734576	0.863	1	6	358.35	3.1	yes	-
136	3330979	0.862	2	5	391.81	2.8	yes	-
137	16728148	0.861	3	9	496.56	2.5	yes	-
138	4925241	0.861	3	7	415.36	2.6	yes	-
139	4925207	0.858	4	5	319.31	0.8	yes	6232205
140	4491946	0.855	2	3	558.23	5.2	no	-
141	4328136	0.855	2	7	517.57	5.5	no	6070722
142	17114061	0.855	2	3	513.78	5	no	17103718
143	9803100	0.854	2	9	436.74	3.8	yes	-
144	17965751	0.851	3	9	447.82	2.9	yes	-
145	1479227	0.845	2	7	435.85	3.3	yes	1479228 3849623
146	10164687	0.844	4	6	514.61	4	no	-
147	17204763	0.843	3	3	399.51	3.5	yes	-
148	17102375	0.843	2	5	357.36	2.2	yes	-
149	1295892	0.838	3	3	476.38	4.5	yes	-
150	1091333	0.837	2	4	318.39	2.6	yes	-
151	10007408	0.836	2	8	538.28	3.4	no	16110566 16110573
152	9931281	0.833	1	6	408.44	3.9	yes	18753351
153	22185388	0.833	2	5	376.24	4.2	yes	-
154	17188031	0.832	3	6	431.42	2.3	yes	-
155	21313984	0.831	5	6	315.27	N.A.	-	-
156	22776612	0.830	3	5	416.45	2.4	yes	-
157	22777710	0.827	2	3	422.33	4.6	yes	-
158	11374997	0.826	4	6	477.96	3.9	yes	-
159	20567489	0.823	2	7	272.33	0.6	yes	-
160	22287166	0.822	3	9	451.8	3.1	yes	-
161	22778059	0.819	3	2	424.34	4.4	yes	-
162	1293604	0.817	2	4	416.2	4	yes	-
163	17099706	0.817	2	4	324.75	3.6	yes	-
164	16738010	0.815	2	8	375.29	3.4	yes	-
165	17204770	0.813	3	3	492.43	4.7	yes	-
166	3334440	0.812	3	3	450.35	3.3	yes	17115571

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
167	3792970	0.811	3	3	405.9	3.1	yes	-
168	21214368	0.806	3	4	465.32	3.7	yes	-
169	21214367	0.806	3	4	420.87	3.5	yes	-
170	3792963	0.804	3	4	389.44	2.6	yes	-
171	17356128	0.804	4	4	521.43	4.2	no	-
172	17094662	0.802	4	4	491.36	3	yes	-
173	4165138	0.802	1	3	377.52	3.7	yes	-
174	24369968	0.798	2	8	456.87	5.2	no	-
175	11314884	0.796	2	5	376.45	3.2	yes	-
176	4052841	0.794	2	4	290.3	2.9	yes	-
177	18870375	0.791	2	5	450.28	3.4	yes	-
178	18870370	0.791	2	5	405.83	3.2	yes	-
179	150171	0.788	4	6	312.82	N.A.	-	-
180	12923393	0.785	5	5	300.25	N.A.	-	-
181	3334436	0.784	3	4	389.44	2.6	yes	-
182	20509949	0.784	3	7	267.35	0.4	yes	-
183	17175825	0.783	4	5	442.49	2.5	yes	-
184	2232090	0.780	2	4	447.31	4.2	yes	-
185	8081443	0.779	1	8	418.41	4.8	yes	-
186	18870385	0.779	3	5	404.84	2.9	yes	-
187	10993963	0.779	1	8	445.43	6.1	no	-
188	4925482	0.778	4	5	429.45	2.5	yes	-
189	21213288	0.775	2	5	357.36	2.6	yes	-
190	17188106	0.774	3	2	410.32	3.8	yes	-
191	15566024	0.772	2	5	288.39	1.9	yes	-
192	7954550	0.770	1	5	360.22	2	yes	-
193	22756470	0.767	3	5	255.34	0.2	yes	-
194	9954936	0.763	2	9	441.38	4.2	yes	23644734
195	20315472	0.754	2	3	263.4	0.9	yes	-
196	24656532	0.746	2	6	404.48	3.1	yes	-
197	960482	0.745	3	3	300.76	2.7	yes	-
198	11386097	0.745	1	6	481.31	5.2	no	-
199	11774947	0.745	1	6	436.86	5.1	no	-
200	960509	0.745	3	3	345.21	2.9	yes	5168147
201	4980279	0.742	2	5	369.41	3.6	yes	9687250
202	21213507	0.741	3	3	529.25	4.5	no	-
203	20242238	0.741	3	7	270.35	N.A.	-	-
204	20242240	0.739	4	7	270.35	N.A.	-	-
205	17322130	0.736	3	3	310.44	1	yes	-
206	24647275	0.733	3	6	420.46	2.8	yes	-
207	24450426	0.732	1	4	382.48	3.8	yes	-
208	18340514	0.730	3	5	407.85	3.3	yes	-
209	20325529	0.730	5	2	303.28	N.A.	-	-
210	16106503	0.729	1	2	363.54	4.2	yes	-
211	17274301	0.728	1	5	528.1	6.1	no	-
212	17117535	0.727	1	5	483.65	5.9	no	-
213	2183698	0.727	1	5	392.2	5.5	no	-
214	9547777	0.726	2	4	336.38	3.1	yes	-
215	8086185	0.725	1	5	419.3	5.1	no	-
216	11730717	0.725	3	10	493.48	3	yes	-
217	256182	0.724	1	3	434.61	5.4	no	3253905
218	20464636	0.724	4	4	275.39	1.5	yes	-
219	17111389	0.723	2	4	473.35	4.5	yes	-
220	21734519	0.720	2	4	333.77	3.1	yes	-
221	8374950	0.719	2	3	301.41	2.4	yes	-
222	10142848	0.718	3	9	522.61	2.6	no	-
223	3008386	0.718	2	5	247.21	1.9	yes	-
224	24883446	0.716	3	7	288.8	N.A.	-	-
225	4924864	0.715	3	5	355.39	3.3	yes	9686049
226	17133656	0.715	3	3	394.25	4.1	yes	-
227	18801044	0.714	1	5	365.42	3.7	yes	-
228	50963	0.714	4	6	288.8	N.A.	-	-
229	24886962	0.714	3	6	333.25	N.A.	-	-
230	21162058	0.710	4	7	292.3	0.7	yes	-
231	9862196	0.709	5	9	327.41	N.A.	-	-
232	19367344	0.709	3	6	276.64	N.A.	-	-
233	22513787	0.708	1	3	402.51	5.5	no	-
234	24855654	0.707	1	5	406.86	4	yes	-
235	20242495	0.704	6	10	586.84	N.A.	no	-
236	21106465	0.704	3	6	416.48	4.3	yes	-
237	20242239	0.702	3	6	287.4	N.A.	-	-
238	20481990	0.702	2	8	321.42	1.1	yes	-
239	10309289	0.700	1	4	374.45	4	yes	-

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
240	1295042	0.698	3	3	371.45	2.5	yes	-
241	9823440	0.695	1	7	420.4	4.6	yes	11154359 18753350
242	17608765	0.694	2	7	375.49	0.7	yes	-
243	7501868	0.692	2	7	406.36	4.8	yes	-
244	595266	0.692	0	0	270.43	6.2	no	-
245	21734488	0.688	1	6	366.37	2.8	yes	-
246	1298267	0.688	1	5	392.2	5.5	no	-
247	21213506	0.684	3	3	440.34	4.1	yes	-
248	7204924	0.683	2	5	358.34	3.9	yes	-
249	19373025	0.683	2	6	266.37	1.4	yes	-
250	1298949	0.680	2	3	381.25	3.5	yes	-
251	4531507	0.678	2	4	369.44	2.6	yes	-
252	16773933	0.677	2	2	212.68	1.6	yes	-
253	7160090	0.672	2	5	381.38	3.7	yes	-
254	4135662	0.671	0	7	521.56	4.7	no	-
255	9870879	0.671	2	8	507.6	2.9	no	-
256	1295682	0.670	3	3	405.9	3.1	yes	-
257	1296020	0.670	3	3	450.35	3.3	yes	17188034
258	20349421	0.668	4	8	307.31	0.6	yes	-
259	22543479	0.668	3	7	313.38	0.5	yes	-
260	23188270	0.662	1	4	432.6	5	no	-
261	20315460	0.658	4	5	400.59	0.5	yes	-
262	18753322	0.657	0	7	434.43	4.7	yes	-
263	9976264	0.657	4	7	367.54	1.6	yes	-
264	22287414	0.657	3	5	414.93	2.5	yes	-
265	13559295	0.654	1	3	443.62	4.8	yes	-
266	20387431	0.652	3	5	305.85	N.A.	-	-
267	3104580	0.647	2	8	533.57	4.8	no	9591740
268	21214340	0.647	3	2	410.32	4.2	yes	-
269	21061429	0.645	3	9	473.52	2.1	yes	-
270	1061578	0.642	1	3	439.3	5	no	-
271	15595237	0.638	3	6	242.3	2.1	yes	-
272	10287860	0.636	2	10	420.29	3.3	yes	-
273	1080673	0.635	2	7	358.34	3	yes	-
274	22879228	0.635	2	2	196.23	0.1	yes	-
275	1297142	0.634	3	3	476.39	3.5	yes	1297141
276	17051758	0.633	3	4	462.34	4.4	yes	-
277	9018511	0.631	1	3	420.52	4.7	yes	-
278	11733093	0.631	1	5	451.57	3.4	yes	21989807
279	20464646	0.631	4	3	245.37	0.4	yes	-
280	21313944	0.631	2	6	257.36	0.4	yes	-
281	10403217	0.630	3	6	311.72	0.6	yes	-
282	18758670	0.628	3	6	512.72	3.4	no	-
283	22898960	0.624	3	7	471.88	N.A.	-	-
284	22898958	0.624	3	7	516.33	N.A.	no	-
285	10612524	0.623	2	6	276.38	0.5	yes	-
286	17189280	0.621	2	3	419.31	4.4	yes	-
287	8075566	0.617	1	5	372.44	2.8	yes	-
288	22287492	0.617	3	7	419.8	4.3	yes	-
289	17312168	0.613	3	5	437.51	4	yes	-
290	20481980	0.611	3	3	274.41	1	yes	-
291	583171	0.611	0	6	370.42	3.4	yes	5371858
292	20481961	0.611	3	3	246.35	2	yes	-
293	23037434	0.611	3	4	372.29	N.A.	-	-
294	18758737	0.611	3	5	482.18	3.6	yes	-
295	4531818	0.610	2	7	474.51	1.9	yes	-
296	21313983	0.610	5	6	316.25	N.A.	-	-
297	21484610	0.607	2	3	187.18	0.2	yes	-
298	21162246	0.604	3	6	398.61	2.7	yes	-
299	12810819	0.601	5	8	312.44	0.2	yes	-
300	20761501	0.600	1	4	468.03	3.6	yes	21989813
301	10108389	0.600	3	7	292.36	1.2	yes	-
302	20761500	0.599	1	4	512.48	3.9	no	21989844
303	23487158	0.599	2	7	294.33	2.2	yes	-
304	17132060	0.599	1	8	613.67	4.9	no	18773024
305	2212652	0.597	3	4	385.44	2.7	yes	-
306	11441627	0.594	3	9	423.34	3.1	yes	-
307	12718697	0.590	3	2	230.35	0.1	yes	-
308	3962581	0.589	3	4	385.44	2.3	yes	-
309	7317223	0.587	0	4	254.7	0.1	yes	-
310	21734493	0.587	1	6	366.37	2.8	yes	-
311	17445779	0.585	2	3	343.85	N.A.	-	-
312	9354415	0.585	0	2	344.2	4.2	yes	-

continued on next page.

continued from previous page.

	ID	prob.	H don.	H accept.	M.W.	xlogp	Lipinski5	compounds of same connectivity
313	16110574	0.583	2	7	471.3	2.4	yes	16110570
314	11741199	0.582	2	7	391.74	3.9	yes	-
315	10096800	0.582	2	7	483.2	4.3	yes	-
316	17074186	0.580	4	4	507.4	3.7	no	-
317	18758696	0.576	3	5	529.18	3.9	no	-
318	18758677	0.576	3	5	526.63	3.8	no	-
319	20507757	0.575	3	7	278.35	1	yes	-
320	2685965	0.573	1	3	310.35	3.1	yes	-
321	24876200	0.573	4	10	569.52	6.7	no	-
322	20342110	0.573	1	5	293.41	1.1	yes	-
323	22756451	0.571	2	7	305.42	0.9	yes	-
324	22983962	0.570	1	2	378.55	4.9	yes	-
325	5126782	0.570	0	6	544.66	6.3	no	-
326	11259048	0.568	3	8	433.38	2.4	yes	-
327	21214375	0.567	3	3	476.38	4.9	yes	-
328	17317146	0.566	3	3	529.25	4.1	no	-
329	720275	0.566	0	3	252.31	3.5	yes	-
330	23439440	0.566	3	11	496.8	3	no	-
331	20875247	0.565	2	4	354.37	4	yes	-
332	20464617	0.564	4	7	272.33	0.4	yes	-
333	5222461	0.563	2	3	339.41	2.7	yes	-
334	1317630	0.563	1	4	388.46	4.4	yes	-
335	24641730	0.563	2	3	395.28	3.5	yes	-
336	1476429	0.560	2	8	405.36	2.8	yes	1476428 3839686
337	22756474	0.559	3	4	286.42	0.3	yes	-
338	22898948	0.559	2	6	380.38	N.A.	-	-
339	4927445	0.558	2	5	376.47	3.1	yes	-
340	20242497	0.558	2	5	307.44	N.A.	-	-
341	3067289	0.555	1	5	229.03	0.5	yes	-
342	17354280	0.555	1	5	184.58	0.4	yes	-
343	4952789	0.555	1	4	480.55	5.9	no	-
344	18758672	0.554	3	5	403.28	2.8	yes	-
345	20464648	0.552	1	9	321.32	2.3	yes	-
346	17231866	0.551	1	3	388.46	4.7	yes	-
347	17646921	0.549	1	5	416.51	3.4	yes	-
348	11383043	0.548	0	5	368.41	3	yes	-
349	20875292	0.546	2	4	354.37	4	yes	-
350	17227780	0.544	1	3	453.33	4.8	yes	-
351	17227758	0.544	1	3	408.88	4.6	yes	-
352	18943116	0.542	2	6	252.34	1	yes	-
353	62949	0.539	3	7	268.34	1	yes	-
354	5098118	0.538	0	4	423.5	5.2	no	-
355	4095823	0.537	1	2	304.43	2.6	yes	-
356	20387430	0.537	3	3	282.86	N.A.	-	-
357	18859966	0.535	2	3	386.87	4.1	yes	-
358	7183435	0.535	0	5	379.76	4.6	yes	-
359	23393282	0.532	1	4	176.18	0.6	yes	-
360	20550125	0.532	5	9	414.51	1.9	yes	-
361	17189325	0.531	3	3	371.45	2.9	yes	-
362	22781186	0.531	3	3	440.34	3.7	yes	-
363	23506021	0.529	2	13	561.45	7.5	no	-
364	17204781	0.524	3	4	508.43	3.7	no	-
365	21624702	0.524	3	5	213.3	1.6	yes	-
366	1476454	0.523	2	4	422.71	3.6	yes	1476453 3343811
367	1476449	0.522	2	4	467.16	3.8	yes	1476450 5134253
368	4162404	0.522	1	3	507.04	5.8	no	-
369	20613565	0.519	2	4	404.93	1.9	yes	-
370	953567	0.517	2	5	346.43	3.1	yes	-
371	4982460	0.516	3	5	416.45	2.4	yes	-
372	7160024	0.516	2	3	370.83	4.4	yes	-
373	21162245	0.515	6	6	508	N.A.	no	-
374	3329649	0.515	2	3	402.89	3.2	yes	-
375	21313936	0.514	3	8	397.56	1.2	yes	-
376	18758667	0.513	3	5	494.73	3.3	yes	-
377	18860030	0.512	2	5	400.44	3.3	yes	-
378	24855685	0.511	1	7	423.81	4.9	yes	-
379	11260539	0.509	3	8	500.27	2.7	no	11318796
380	4648082	0.509	0	10	575.55	5.4	no	-
381	11775427	0.509	3	8	455.81	2.3	yes	-
382	21966558	0.508	1	3	269.3	2.7	yes	-
383	17429343	0.506	2	3	317.41	1.7	yes	-
384	20242498	0.500	1	6	306.43	N.A.	-	-

1: PubChem Compound ID, 2: predicted probability, 3: number of H bond donor, 4: number of H bond acceptor, 5: molecular weight, 6: whether a chemical compound satisfies Lipinski's rule of five (# of H bond donor $\leq$ 5 & # of H bond acceptor $\leq$ 10 & molecular weight $\leq$ 500 & xlogp $\leq$ 5), 7: not available because some values were not available.

Appendix B - Supplementary explanation for SVM

Basic theory

The Support Vector Machine (SVM) is a statistical learning algorithm for binary classification on the basis of the following theory (Cortes and Vapnik, 1995; Vapnik, 1998; Cristianini and Sawe-Taylor, 2000).

Linear SVM; the linearly separable case

Given l samples which belong to one of the two classes, χ_1 and χ_2, and are represented by n-dimensional vectors vectors,

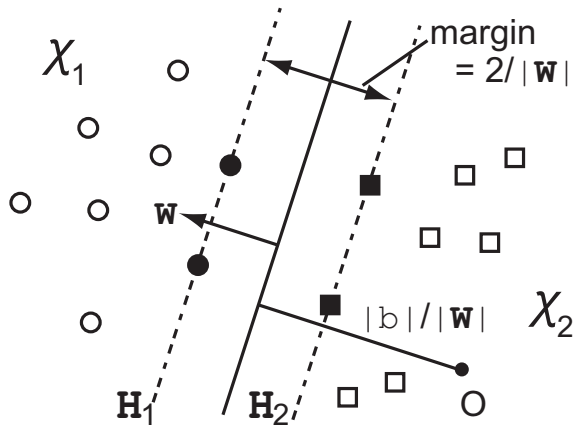
$$\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_l, y_l)\} \quad \boldsymbol{x}_i \in \mathrm{R}^n, \; y_i \in \{+1, -1\}$$
$$y_i = \begin{cases} 1 & \boldsymbol{x}_i \in \chi_1 \\ -1 & \boldsymbol{x}_i \in \chi_2 \end{cases} \tag{B1}$$

the SVM produces the following classification rule $f_{\boldsymbol{w}}$ that decides where a sample $\boldsymbol{x}$ belongs to χ_1 or χ_2,

$$f_{\boldsymbol{w}}(\boldsymbol{x}) \equiv \mathrm{sign}(g(\boldsymbol{x})) = \mathrm{sign}(\boldsymbol{w}^t \boldsymbol{x} + b) = \begin{cases} 1 & \boldsymbol{x} \in \chi_1 \\ -1 & \boldsymbol{x} \in \chi_2 \end{cases} \tag{B2}$$

where sign function expresses

$$\mathrm{sign}(g(\boldsymbol{x})) \equiv \begin{cases} 1 & g(\boldsymbol{x}) \geq 0 \\ -1 & g(\boldsymbol{x}) < 0 \end{cases}.$$



Supplementary Fig. B1 Linear SVM applied to linearly separable cases

Eq. (B2) is correspond to the identification of a boundary including a boundary line, boundary plane or boundary hyperplane that separates the two classes given samples belonging to one of the two classes and represented in the n-dimensional space by a feature vector $\boldsymbol{x}$ (Fig. B1). On the boundary, $g(\boldsymbol{x}) = \boldsymbol{w}^t\boldsymbol{x} + b$ in Eq. (B2) equals to 0.

In the linearly separable case, there exist several boundaries that correctly classify all the samples. For example, both a solid line and two dotted lines in Fig. B1 can be a boundary.

From these several boundaries, the SVM searches for the boundary that maximizes the margin (Fig. B1) as the most favorable boundary. The margin is a area around the boundary where no training samples exist (area between two dotted lines in Fig. B1. The larger the margin, the more powerful the boundary in classifying unknown samples that are distant from training or known samples and difficult to classify. Because the samples that are not easy to discriminate are likely to appear around the boundary, or in the margin area, the large margin facilitates the classification of these samples into the correct classes.

Taking the margin into consideration, $g(\boldsymbol{x})$ in Eq. (B1) can be defined as follows.

$$\forall i, \quad g(\boldsymbol{x}) = \boldsymbol{w}^t\boldsymbol{x} + b \begin{cases} \geq 1 & \text{if } (y_i \in 1) \\ \leq -1 & \text{if } (y_i \in -1) \end{cases} \tag{B3}$$

Eq. (B3) shows that there exist no training samples in the area between two lines (planes or hyperplanes) $g(\boldsymbol{x}) = \pm 1$ (H_1 and H_2 in Fig. B1). The distance between H_1 and H_2 is the margin under this condition.

Eq. (B3) can be rewritten with y in Eq.(B1) as follows.

$$y_i \cdot (\boldsymbol{w}^t\boldsymbol{x}_i + b) \geq 1 \quad (i = 1, \ldots, l) \tag{B4}$$

As the distance between any sample $\boldsymbol{x}$ and the boundary $g(\boldsymbol{x}) = 0$ is $|g(\boldsymbol{x})|/\|\boldsymbol{w}\|$, the margin, or distance, between H_1 and H_2 where $g(\boldsymbol{x}) = \pm 1$ is $2/\|\boldsymbol{w}\|$. Therefore, the maximization of the margin, which is the goal of the SVM, can be expressed as follows.

$$\begin{aligned} &\underset{\boldsymbol{w},b}{\text{Minimize}} \qquad G(\boldsymbol{w}) = \tfrac{1}{2}\|\boldsymbol{w}\|^2 \\ &\text{subject to} \quad \forall i, \quad y_i(\boldsymbol{w}^t\boldsymbol{x}_i + b) - 1 \geq 0 \quad (i = 1, \ldots, l) \end{aligned} \tag{B5}$$

This minimization problem can be solved by using Lagrange multipliers. With $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_l)$ where α_i is a positive variable, Lagrange function L is as follows.

$$L(\boldsymbol{w}, b, \boldsymbol{\alpha}) \equiv \tfrac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{l} \alpha_i [y_i \cdot (\boldsymbol{w}^t\boldsymbol{x}_i + b) - 1] \tag{B6}$$

The following partial differentiations of L are solved to identify w and b.

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\alpha})}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{l} \alpha_i y_i \boldsymbol{x}_i = 0 \tag{B7}$$

$$\frac{\partial L(\boldsymbol{w}, b, \boldsymbol{\alpha})}{\partial b} = -\sum_{i=1}^{l} \alpha_i y_i \boldsymbol{x}_i = -\boldsymbol{\alpha}^t \boldsymbol{y} = 0 \tag{B8}$$

On the basis of Eq. (B7), $\boldsymbol{w}^*$, a solution for the minimization problem, is obtained as follows.

$$\boldsymbol{w}^* = \sum_{i=1}^{l} \alpha_i y_i \boldsymbol{x}_i$$

Substitution of this $\boldsymbol{w}^*$ into Eq. (B6) and Eq. B8 produce

$$\begin{aligned} L(\boldsymbol{\alpha}) &= \sum_{i=1}^{l} \alpha_i - \frac{1}{2}|\boldsymbol{w}^*|^2 \\ &= \boldsymbol{e}^t \boldsymbol{\alpha} - \tfrac{1}{2}\boldsymbol{\alpha}^t Q \boldsymbol{\alpha} \end{aligned} \tag{B9}$$

where $\boldsymbol{e} = (1, 1, \cdots, 1)^t$, $Q_{ij} = y_i y_j \boldsymbol{x}_i^t \boldsymbol{x}_j$.

Therefore, the minimization problem Eq. (B5) can be converted to the following maximization problem.

$$\begin{aligned} &\underset{\boldsymbol{\alpha}}{\text{Maximize}} \quad L(\boldsymbol{\alpha}) = \boldsymbol{e}^t \boldsymbol{\alpha} - \tfrac{1}{2}\boldsymbol{\alpha}^t Q \boldsymbol{\alpha} \\ &\text{subject to} \quad \boldsymbol{\alpha}^t \boldsymbol{y} = 0, \quad \alpha_i \geq 0 \quad (i = 1, \cdots, l) \end{aligned} \tag{B10}$$

Eq. B10) is a dual problem for Eq. (B5) in the quadratic programming. $\boldsymbol{w}^*$ is derived only from samples $\boldsymbol{x}_i$ whose corresponding $\alpha_i^* > 0$, which is the solution for Eq. B10). These samples constructing the boundary are called support vectors (SV). In Fig. B1, training samples on the H_1 and H_2 are SVs.

As for b^*, on the basis of the complementary condition in nonlinear programming

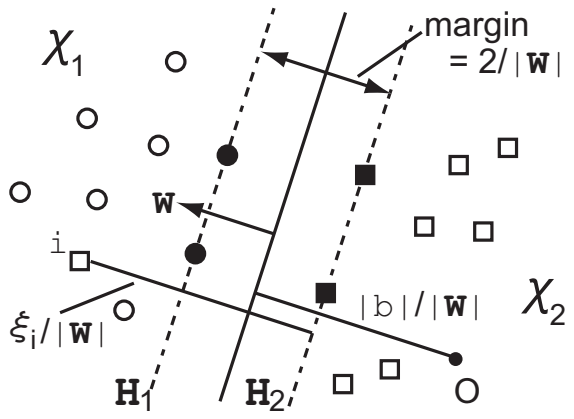
$$\forall i, \quad \boldsymbol{\alpha}_i^* [y_i \cdot (\boldsymbol{w}^{*t} \boldsymbol{x}_i + b^*) - 1] = 0, \tag{B11}$$

b^* is decided by using any SV $\boldsymbol{x}_i$ ($\alpha_i^* > 0$) as follows.

$$b^* = y_i - \boldsymbol{w}^{*t} \boldsymbol{x}_i \tag{B12}$$

In conclusion, substitution of Eq.—(IV) and Eq.—(B12) into Eq.—(B2) produces the classification function of linear SVM

$$f(\boldsymbol{x}) = \text{sign}\left(\sum_{i; x_i \in SVs} \alpha_i y_i \boldsymbol{x_i}^t \boldsymbol{x} + b^* \right). \tag{B13}$$



Supplementary Fig. B2 Linear SVM applied to linearly unseparable cases

Linear SVM; the linearly unseparable case

In the linearly unseparable case, $\boldsymbol{w}$ satisfying Eq. (B3) doesn't exist. Thus, the condition of the margin is relaxed by introducing the variable $\boldsymbol{\xi}$ ($\forall i,\ \xi_i \geq 0,\quad i = 0, \ldots, l$) or permitting misclassification. For example, in Fig. B2, the sample i is in the margin area and $\xi_i > 0$.

Taking $\boldsymbol{\xi}$ into consideration, Eq. (B3) is converted into

$$\boldsymbol{w}^t \boldsymbol{x} + b \begin{cases} \geq 1 - \xi_i & \text{if } \boldsymbol{x}_i \in \chi_1) \\ \leq -1 + \xi_i & \text{if } \boldsymbol{x}_i \in \chi_2) \end{cases}. \tag{B14}$$

Eq. (B5) is also converted into

$$\begin{aligned} \underset{\boldsymbol{w},b,\boldsymbol{\xi}}{\text{Minimize}} \quad & G(\boldsymbol{w}, \boldsymbol{\xi}) = \tfrac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{i=1}^{l} \xi_i \\ \text{subject to} \quad & \forall i, y_i(\boldsymbol{w}^t \boldsymbol{x}_i + b) - (1 - \xi_i) \geq 0 \quad (i = 1, \ldots, l) \\ & \forall i, \xi_i \geq 0 \end{aligned} \tag{B15}$$

where the second term is the penalty for samples that lie in the margin. The summation of ξ_i gives an upper limit of the number of the misclassified training samples. The constant C decides the balance between the first and the second term. As the lower C permits the larger ξ, the distance between $g(\boldsymbol{x}) = \pm 1$ (two dotted lines in Fig. B2) gets larger.

Eq. (B15) can be solved by using Lagrange multipliers as follows.

$$\begin{aligned} L(y, \boldsymbol{x}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \equiv \quad & \tfrac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{l} \alpha_i [y_i(\boldsymbol{w}^t \boldsymbol{x}_i + b) - 1 + \xi_i] \\ & - \sum_{i=1}^{l} \gamma_i \xi_i + C \sum_{i=1}^{l} \xi_i \end{aligned} \tag{B16}$$

Partial differentiation of L gives

$$\frac{\partial L(y, \boldsymbol{x}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma})}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{l} \alpha_i y_i \boldsymbol{x}_i = 0$$

$$\frac{\partial L(y, \boldsymbol{x}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma})}{\partial b} = \boldsymbol{\alpha}^t \boldsymbol{y} = 0$$

$$\frac{\partial L(y, \boldsymbol{x}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma})}{\partial \boldsymbol{\xi}} = C - \alpha_i - \gamma_i = 0.$$

On the basis of these equations, the dual problem for Eq. (B15), which corresponds to Eq. (B10), is

$$\begin{aligned} & \underset{\boldsymbol{\alpha}}{\text{Minimize}} \quad \tfrac{1}{2}\boldsymbol{\alpha}^t Q \boldsymbol{\alpha} - \boldsymbol{e}^t \boldsymbol{\alpha} \\ & \text{subject to} \quad \boldsymbol{\alpha}^t \boldsymbol{y} = 0, \quad 0 \le \alpha_i \le C \quad (i = 1, \cdots, l). \end{aligned} \tag{B17}$$

On the hand, based on the complementary condition, the following equations can be derived.

$$\begin{aligned} \forall i, \quad & \boldsymbol{\alpha}_i^* [y_i \cdot (\boldsymbol{w}^{*t} \boldsymbol{x}_i + b^*) - 1 + \xi_i] = 0 \\ & \forall i, \quad \gamma_i \xi_i = 0 \end{aligned} \tag{B18}$$

b^*, which is the solution o Eq. (B15) can be derived on the basis of Eq. (B18) by using any $\boldsymbol{x}_i$ satisfying $0 < \alpha_i^* < C$.

In conclusion, the classification function of the linear SVM in the linearly unseparable case is the same as that in the linearly separable case and is

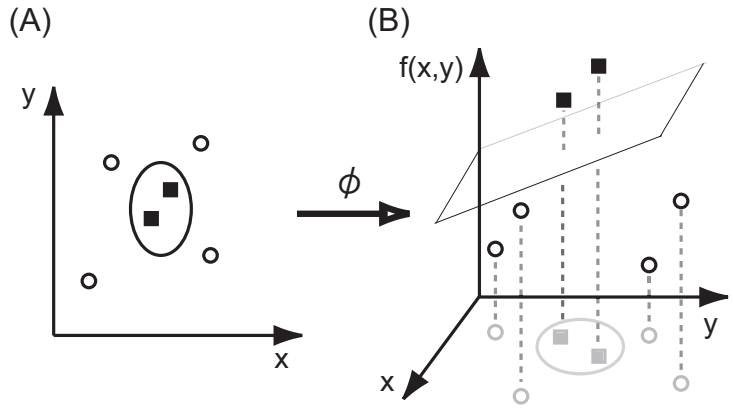
$$f(\boldsymbol{x}) = \text{sign}\left(\sum_{i; x_i \in SVs} \alpha_i y_i \boldsymbol{x_i}^t \boldsymbol{x} + b^* \right)$$

Here, $\boldsymbol{x}_i$ $(i = 1, \cdots, l)$ is classified into the following three cases.

(1) $g(\boldsymbol{x}_i)/y_i > 1,\ \alpha_i = 0,\ \gamma_i = C, \xi = 0$

(2) $g(\boldsymbol{x}_i)/y_i = y_i,\ 0 < \alpha_i < C,\ 0 < \gamma_i < C,\ \xi = 0$

(3) $g(\boldsymbol{x}_i)/y_i < 1,\ \alpha_i = C,\ \gamma_i = 0,\ \xi \neq 0$



Supplementary Fig. B3 Kernel functions

In (2) and (3), $bvecx_i$ is SV. In (3), $\boldsymbol{x}_i$ is lies between H_1 and H_2 in Fig. B2. When $\boldsymbol{x}_i \in \chi_i$, it is correctly classified if it lies in between H_1 and the boundary, and is misclassified if it is between the boundary and H_2.

Non-linear SVM; Kernel trick

When the boundary can't be approximated by a hyperplane and there exists a complex classification boundary (Fig. B3A), the classification function should be a non-linear function in order to achieve high classification performances.

However, by considering a function ϕ, which is a mapping to the higher-dimension,

$$\phi(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), \ldots, \phi_n(\boldsymbol{x}), \ldots, \phi_{n'}(\boldsymbol{x}))^t, \tag{B19}$$

samples belonging to the two classes can be classified in the high dimensional space $\phi(\boldsymbol{x})$ (Fig. B3B). Here, the linear classification boundary in the converted $\phi(\boldsymbol{x})$ space constitutes non-linear classification boundary in $\boldsymbol{x}$ feature space.

In Eq. (B10) and Eq. (B17) by replacing $Q_{ij} = y_i y_j \boldsymbol{x}_i^t \boldsymbol{x}_j$ with

$$Q_{ij} = y_i y_j \phi(\boldsymbol{x}_i)^t \phi(\boldsymbol{x}_j), \tag{B20}$$

the classification function

$$f(\boldsymbol{x}) = \text{sign}\left(\sum_{i; x_i \in SVs} \alpha_i y_i \phi(\boldsymbol{x_i})^t \phi(\boldsymbol{x}) + b^* \right) \tag{B21}$$

is obtained.

In Eq. (B21), calculation of $\phi(\boldsymbol{x})$ is generally time consuming and can be infeasible. On the other hand, $\phi(\boldsymbol{x})$ always appears as inner product. Therefore, only if inner product of ϕ can be defined, it is not necessary to calculate $\phi(\boldsymbol{x})$.

Here, a Kernel function K satisfies

$$K(\boldsymbol{x}, \boldsymbol{y}) \equiv \phi(\boldsymbol{x})^t \phi(\boldsymbol{y}) = \sum_{i=1}^{n'} \phi_i(\boldsymbol{x}) \phi_i(\boldsymbol{y}). \tag{B22}$$

A function K is a Kernel function when it satisfies

$$\int \int K(x, y) g(x) g(y) dx dy \geq 0$$

where x and y are continuous, or when, given input data $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ which can be disperse,

$$Q = (K(\boldsymbol{x}_i . \boldsymbol{x}_j))_{i,j=1}^n$$

is semi-definite, which means that all of the eigenvalues of matrix Q are not negative. The Kernel function include the following functions.

$$\begin{aligned} &\text{(polynomial)} \quad K(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\gamma \boldsymbol{x}_i^t \boldsymbol{x}_j + r)^d \\ &\text{(RBF)} \qquad\quad K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2) \\ &\text{(sigmoid)} \qquad K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \tanh(\gamma \boldsymbol{x}_i^t \boldsymbol{x}_j + r) \end{aligned} \tag{B23}$$

By using the Kernel function, Eq. (B20) is converted into

$$Q_{ij} = y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j).$$

In conclusion, the classification function is

$$f(\boldsymbol{x}) = \text{sign} \left(\sum_{i; x_i \in SVs} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b^* \right). \tag{B24}$$

Probability output of SVM

The output of SVM can be treated as probability according to Platt, 2000.
Given

$$\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_l, y_l)\} \quad \boldsymbol{x}_i \in \mathrm{R}^n, \; y_i \in \{+1, -1\}$$

$$y_i = \begin{cases} 1 & \boldsymbol{x}_i \in \chi_1 \\ -1 & \boldsymbol{x}_i \in \chi_2 \end{cases},$$

the probability is obtained by

$$p(y = 1 | \boldsymbol{x}) = \frac{1}{1 + \exp(A f(\boldsymbol{x}) + B)}$$

where

$$f(\boldsymbol{x}) = \sum_{i;x_i \in SVs} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b.^*$$

Parameters A and B is obtained by solving the maximum likelihood estimation

$$\min_{z=(A,B)} -\sum_{i=1}^{l} (t_i \log p_i) + (1 - t_i) \log(1 - p_i))$$

where

$$p_i = \frac{1}{1 + \exp(Af(\boldsymbol{x}_i) + B)}, \quad t_i = \begin{cases} \frac{N_+ + 1}{N_+ + 2} & \text{if } \ y_i = 1 \\ \frac{1}{N_- + 2} & \text{if } y_i = -1 \end{cases}, \quad i = 1, 2, \ldots, l$$

and training data consist of N_+ positive or $y_i = 1$ samples and N_- negative or $y_i \ -1$ samples.