Acta Pharmacologica Sinica (2014) 35: 419–431
 1671-4083/14 $32.00
www.nature.com/aps

Original Article

Multi-algorithm and multi-model based drug target prediction and web server

Ying-tao LIU[1, #], Yi LI[1, #], Zi-fu HUANG[1, #], Zhi-jian XU[1], Zhuo YANG[1], Zhu-xi CHEN[1], Kai-xian CHEN[1], Ji-ye SHI[2, *], Wei-liang ZHU[1, *]

[1]*Drug Discovery and Design Center, Key Laboratory of Receptor Research, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai 201203, China;* [2]*Informatics Department, UCB Pharma, 216 Bath Road, Slough SL1 4EN, UK*

Aim: To develop a reliable computational approach for predicting potential drug targets based merely on protein sequence.
Methods: With drug target and non-target datasets prepared and 3 classification algorithms (Support Vector Machine, Neural Network and Decision Tree), a multi-algorithm and multi-model based strategy was employed for constructing models to predict potential drug targets.
Results: Twenty one prediction models for each of the 3 algorithms were successfully developed. Our evaluation results showed that ~30% of human proteins were potential drug targets, and ~40% of putative targets for the drugs undergoing phase II clinical trials were probably non-targets. A public web server named D3TPredictor (http://www.d3pharma.com/d3tpredictor) was constructed to provide easy access.
Conclusion: Reliable and robust drug target prediction based on protein sequences is achieved using the multi-algorithm and multi-model strategy.

Keywords: drug target; protein sequence; multi-algorithm and multi-model strategy; web server; support vector machine; neural network; decision tree

Acta Pharmacologica Sinica (2014) 35: 419–431; doi: 10.1038/aps.2013.153; published online 3 Feb 2013

Introduction

Drug target identification and validation is typically the first step in the drug discovery process[1]. The estimated number of drug targets in the human proteome ranges from nearly 3000 to more than 10 000[2–4]. However, the number of drug targets validated by marketed drugs is very small in comparison. Drews identified 483 drug targets[5], and a more recent report showed that oral small-molecule drugs target only 186 human targets[6], indicating quite an unsettled circumstance about how many potential drug targets there are in the human proteome. Furthermore, 30%–40% of experimental drugs fail during the drug discovery process because of inappropriate target choice[7]. Therefore, the development of reliable computational approaches for the prediction of new drug targets is extremely valuable.

A number of strategies have been reported for predicting potential drug targets using protein structures or sequences as input[2, 8–16]. The strategies can be generally classified into three groups. The first group nominates new drug targets based on their similarity to known drug targets at the sequence, function and/or domain level[2, 9]. The second group searches for potential binding pockets on the protein surface based on three dimensional (3D) structures and evaluates the druggability of those pockets based on properties such as geometric and energetic features[8, 11]. The third group uses machine-learning algorithms to classify drug targets and non-targets based on descriptors representing biochemical and physicochemical features of proteins[15, 16]. The methods, which are based on machine-learning algorithms such as Support Vector Machine (SVM), Neural Network (NN) and Decision Tree (DT), have been validated to be effective for drug target prediction according to published studies[15, 17–22].

However, the abovementioned groups of methods have their own limitations: the first group is less effective when proteins exert no or low homology to known drug targets; the second group is constrained by the availability of experimentally determined 3D structures; the third group's performance is highly dependent on the quality and quantity of the training

# The first three authors contributed equally to this work.
* To whom correspondence should be addressed.
E-mail Jiye.Shi@ucb.com (Ji-ye SHI)
wlzhu@mail.shcnc.ac.cn (Wei-liang ZHU)
Received 2013-07-20 Accepted 2013-09-23

data. In particular, non-target datasets need to be carefully verified because many were built by simply removing target proteins from protein databases. To the best of our knowledge, no multi-algorithm and multi-model approach for predicting potential drug targets has been reported to date.

In this study, we carefully prepared the target and non-target datasets and employed three machine-learning algorithms, SVM, NN, and DT, to build multiple sequence-based models for the prediction of drug targets. All models were subsequently cross-validated and compared with one other. Based on those results, a multi-algorithm and multi-model strategy was established to provide reliable drug target prediction using only protein sequence information as input. The method has been implemented as a public web server, D^3TPredictor, accessible at http://www.d3pharma.com/d3tpredictor. Therefore, this study provides the scientific community with a new tool and access to drug target prediction.

Materials and methods

Target dataset preparation

The targets of 183 marketed drugs[6] and 172 drug candidates currently in phase III clinical trials were extracted from Swiss-Prot[23] and the Thomson Pharma database[24], respectively. After the elimination of identical entries as well as a target protein of extreme length (22152 amino acids), the remaining targets were retained as the target dataset (T-Set) (Figure 1B).

Non-target dataset preparation

The non-target dataset (NT-Set) needs to be rationally curated and filtered because it is inherently difficult to define a protein as not being a drug target. This step is critical because the quality of the non-target dataset greatly affects the reliability of mathematical models constructed by the machine-learning techniques trained on the target and non-target datasets. The non-target proteins were selected from two sources, the Drug Adverse Reaction Target Database (DART)[25] and the Protein Data Bank (PDB)[26], based on the following steps (Figure 1A).

Filtration 1

Among the 86 proteins from the DART, only those associated with serious side effects, such as carcinogenesis, teratogenesis, neurotoxicity and cardiotoxicity, *etc*, were selected as true non-targets for use in Dataset 1 (size: 46) because significant clinical adverse effects strongly indicate that the proteins are not suitable drug targets.

Filtration 2

The DBREF records in a PDB file provide cross-reference links to external databases (*eg*, GenBank, UNIPROT and Norine). Among the 973 human protein entries in the PDB, only those with unique UNIPROT accession codes were retained to comprise Dataset 2 (size: 400); the cross-reference to UNIPROT[27] was important because it offered aggregate knowledge of each protein and greatly assisted the extraction of binding site/pocket information for use in Filtration 4.

Filtration 3

Each entry in Dataset 2 was used to search against the T-Set using BLAST[28]. Entries with e-value lower than 0.001 to any sequence in the T-Set were considered homologous to a drug target and thus removed. The remaining entries comprised Dataset 3 (size: 194).

Filtration 4

To further remove potential drug targets from Dataset 3, a binding pocket-based SVM model (pbSVM, Figure 1C) was developed based on T-Set (drug targets; Figure 1B) and Dataset 1 (non-targets; Figure 1A). Since the binding pocket char-

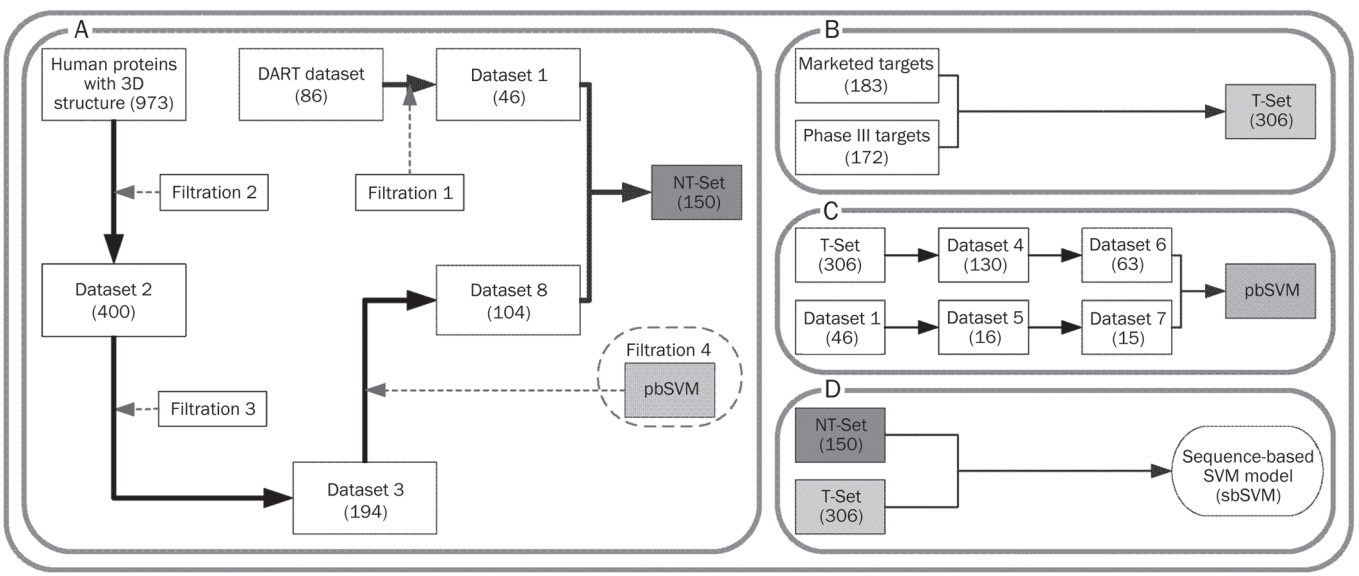

Figure 1. Dataset preparation flowchart.

acterization requires structural data, only entries with known structures from T-Set and Dataset 1 were selected to form Dataset 4 (size: 130) and Dataset 5 (size: 16), respectively.

The binding pockets of proteins in Dataset 4 and Dataset 5 were assigned based on literature search. In cases where there was insufficient or conflicting literature information regarding a binding pocket, the corresponding protein was rejected. Proteins with large, flat binding sites and those with poor-quality structures that hindered accurate characterization of the binding pockets were also rejected. The surviving proteins comprised Dataset 6 (drug targets, size: 63) and Dataset 7 (non-targets, size: 15), respectively.

SYBYL[29] was used to calculate properties for each binding pocket in Dataset 6 and Dataset 7, including surface area, volume, depth, flexibility, hydrophobicity, electrostatic potential and hydrogen bonding sites, producing 12 values for each pocket. Those values, related to the binding affinity between a target and its ligand[12, 14, 30], were normalized according to Equation 1 and used as descriptors to build the pbSVM model using the LIBSVM software package[31] with 5-fold cross-validation (Figure 1C).

$$scaledValue = \frac{oriValue - minValue}{maxValue - minValue} \qquad \text{(Eq 1)}$$

Equation 1 represents the normalization method, where *scaledValue* is the scaled value of a given binding pocket property; *oriValue* is the original value of the given property; *minValue* and *maxValue* are the minimum and the maximum value, respectively, of the given property across both Dataset 6 and Dataset 7.

The resulting binding pocket-based SVM model, pbSVM, was used to detect potential drug targets in Dataset 3. The binding pocket descriptors of each protein in the dataset, whose binding pockets could be identified through literature search and structurally characterized, were calculated and normalized according to the aforementioned approach. Those descriptors were then fed into the pbSVM model, and the protein was classified as either a drug target or a non-target. All entries classified as drug targets were used to search against Dataset 3 using BLAST; any hit with e-value less than 0.001 was deemed a potential drug target and removed from Dataset 3. The remaining entries formed Dataset 8 (size: 104, Figure 1A).

Non-target Dataset

The high-quality non-target dataset (NT-Set) was obtained by combining Dataset 1, derived from the DART, and Dataset 8, derived from the PDB.

Descriptor extraction and selection

To build a reliable sequence-based model, a comprehensive group of 175 physicochemical features (Table 1) was used to represent protein sequences. The features were calculated using our in-house programs as well as free and/or open source tools for academic use[14, 15, 32–37]. Each descriptor was normalized into the range of 0–1 using Equation 1. The 175 normalized descriptors were assembled into a descriptor vector d^{175}.

Table 1. Protein sequence based descriptors.

Dimension	Properties	References
1	Hydrophobicity	37
1	pI value	32
1	protein length	32
1	PEST region	32
1	O-glycosylation number	35
1	N-glycosylation number	34
1	Transmembrane helices number	32
1	Signal peptide cleavage	33
20	Composition of 20 amino acid residues	15
21	Attribute composition	36, 37
21	Attribute transition	36, 37
105	Attribute distribution	36, 37

Since appropriate combinations of descriptors usually result in better performance for machine learning techniques[14, 38], two descriptor selection methods were utilized to search for such combinations.

Randomized descriptor selection

Let d^t be a subset of d^{175} comprising t descriptors randomly selected from the 175 normalized descriptors, where t=100, 105, 110, ···, 170. For each t, 10 d^t vectors were randomly generated, resulting in a total of 150 descriptor vectors. Each of the 150 descriptor vectors, as well as d^{175}, was used to train a model. The 151 models were evaluated and the descriptor vector producing the best-performing model was retained.

F-score based descriptor selection

F-score is a simple, intuitive method used to evaluate the discriminative power of a descriptor. Given a descriptor vector d, if there are N^p positive instances (true positives) and N^n negative instances (true negatives), the *F*-score of the ***i***th descriptor $F(i)$ is calculated as

$$F(i) \equiv \frac{(\bar{d}_i^{\,P} - \bar{d}_i)^2 + (\bar{d}_i^{\,n} - \bar{d}_i)^2}{\frac{1}{N^p - 1}\sum_{k=1}^{N^p}(d_{k,i}^{\,p} - \bar{d}_i^{\,p})^2 + \frac{1}{N^n - 1}\sum_{k=1}^{N^n}(d_{k,i}^{\,n} - \bar{d}_i^{\,n})^2} \qquad \text{(Eq 2)}$$

where $\bar{d}_i$, $\bar{d}_i^{\,P}$, and $\bar{d}_i^{\,n}$ are the average values of the ith descriptor of the entire, the positive, and the negative instances, respectively, and $d_{k,i}^{\,P}$ and $d_{k,i}^{\,n}$ the values of the ith descriptor of the kth positive and negative instance, respectively. The larger the *F*-score is, the more discriminative the descriptor is statistically. Before modeling, the *F*-score of each descriptor, based on the training set, was calculated using Equation 2, and the descriptors were sorted by their *F*-scores in descending order. Let d^p be a subset of d^{175} comprising the top p% of the 175 normalized descriptors, where p=10, 20, ···, 90. For each p, 10 d^p vectors were generated, yielding a total of 90 descriptor vectors. Each of the 90 descriptor vectors, as well as d^{175}, was used to train the models. The 91 models were evaluated, and the

combination of descriptors that produced the best-performing model was retained.

Modeling strategy

Modeling strategy I

1) 120 non-targets from the NT-Set were randomly selected as the negative training dataset; 2) 120 targets were randomly selected from the T-Set as the positive training dataset; 3) the remaining entries in the T-Set (186) and the NT-Set (30) were used as the validation set; 4) three kernel functions were independently used to build SVM models using LIBSVM; 5) 10-fold cross-validation was applied; 6) each modeling procedure was repeated 10 times.

Modeling Strategy II

1) All 150 non-targets in the NT-Set were selected as the negative dataset, and 150 targets were randomly selected from the T-Set as the positive dataset; 2) 100 or 120 entries were randomly selected from each of the positive and the negative datasets as the training set, and the remaining entries in the two datasets served as the validation set; 3) the aforementioned two descriptor selection methods, randomized and F-score based, were utilized independently to search for the best combination of descriptors; 4) three classification algorithms were independently implemented and tested to search for the best modeling parameters; 5) 10-fold cross-validation was applied; 6) each modeling procedure was repeated 10 times.

Performance evaluation

The performance of a model was assessed by sensitivity (Equation 3), specificity (Equation 4) and accuracy (Equation 5).

$$\text{Sensitivity}=\frac{TP}{TP+FN} \quad \text{(Eq 3)}$$

$$\text{Specificity}=\frac{TN}{TP+FP} \quad \text{(Eq 4)}$$

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN} \quad \text{(Eq 5)}$$

TP, *TN*, *FP*, and *FN* represent true positives, true negatives, false positives and false negatives, respectively.

$$\text{ASE}=\sum_{\substack{i=1,2,3,\\ j=1,2,\cdots,21}} \text{SD}(\text{Model}_{ij}^{\text{SVM}}, \text{Model}_{ij}^{\text{NN}}, \text{Model}_{ij}^{\text{DT}}) \quad \text{(Eq 6)}$$

Accumulated standard error (ASE) evaluation

Here *i* stands for Dataset I, II or III; *j* is the model serial of 21 SVM models; $\text{Model}_{ij}^{\text{SVM}}$ denotes the accuracy of the SVM model *j* over Dataset *i*; $\text{Model}_{ij}^{\text{NN}}$ denotes the accuracy of the NN model *j* over Dataset *i*; $\text{Model}_{ij}^{\text{DT}}$ denotes the accuracy of the DT model *j* over Dataset *i*; SD represents standard error among three parallel models over an identical dataset.

Multi-algorithm and/or multi-model based strategy

Multi-algorithm based strategy

A query protein sequence is submitted to *M***N* (*M*=1, 2, ···, 21; *N*=1, 2, 3) models. *M* represents the number of selected training sets, and *N* represents the number of selected algorithms (SVM, NT, DT). For each training set, *N* models are constructed based on *N* algorithms, and those *N* models are called "parallel models". *M***N* models are used to classify the query sequence, yielding *M***N* labels (target or non-target) for the query sequence. Then, every *N* labels based on an identical training set are reduced to one label, the one observed in the majority of the *N* labels. Hence, *M***N* labels are reduced to *M* labels, which is called a multi-algorithm based strategy.

Multi-model based strategy

A query sequence is submitted to *M***N* (*M*=1, 2, ···, 21; *N*=1, 2, 3) models, yielding *M***N* labels. Then, every *M* labels using the same algorithm are reduced to one label, the one observed in the majority of the *M* labels. Hence, *M***N* labels are reduced to *N* labels, which is called a multi-model based strategy.

Multi-algorithm and multi-model based strategy

A query sequence is submitted to *M***N* (*M*=1, 2, ···, or 21; *N*=1, 2, or 3) models, yielding *M***N* labels. First, the multi-algorithm based strategy is implemented to reduce *M***N* labels to *M* labels; subsequently, the multi-model based strategy is implemented to reduce *M* labels to one label. This is called a multi-algorithm and multi-model based strategy.

Results

Target dataset

T-Set, totaling 306 entries, was prepared from targets of marketed drugs and drug candidates in phase III clinical trials through elimination of redundancy and manual selection (Figure 1B; see Materials and Methods).

Non-target dataset

NT-Set was prepared from 86 potential non-target proteins in the DART[25] and 973 human proteins in the PDB[26] through 4 filtration steps (Figure 1A, 1C; see Materials an methods).

First, 46 non-target proteins (Dataset 1) were extracted from the DART through Filtration 1. Then, sequence-similarity based filtrations (Filtrations 2 and 3) extracted 194 (Dataset 3) potential non-targets from the 973 human proteins with 3D structures. In Filtration 4, the pbSVM model, whose cross-validation accuracy and testing accuracy were 86.7% and 92.3%, respectively, predicted that there were 104 non-targets (Dataset 8) in Dataset 3. Finally, Datasets 1 and 8 were combined to form the NT-Set of 150 non-target proteins. This set, together with the T-Set of 306 drug targets, was utilized to construct sequence-based models for drug target prediction.

Classification algorithms and selection of SVM kernel function

SVM, NN, and DT are three widely used classification algorithms, each with unique advantages and disadvantages. To the best of our knowledge, most studies on drug target prediction employed only a single algorithm[11, 14, 17, 19, 20, 22]. To explore potential synergies, all three algorithms were implemented and compared in this study.

Three kernel functions, linear, polynomial and radial basis function (RBF), are commonly used in SVM methods. To identify the optimal parameters and the best-performing SVM model, all three kernel functions were evaluated according to Modeling Strategy I (Materials and methods). For the linear kernel function (Equation 7), only one parameter, the error trade-off *C*, was optimized; for the polynomial kernel function (Equation 8), two kernel parameters, γ and d, were optimized; and for the RBF (Equation 9), kernel parameter γ and the error trade-off *C* were optimized.

$$K(x_i, x_j)=x_i^T \cdot x_j \quad \text{(Eq 7)}$$

$$K(x_i, x_j)=(\gamma x_i^T \cdot x_j+1)^d, \gamma>0 \quad \text{(Eq 8)}$$

$$K(x_i, x_j)=exp(-\gamma \parallel x_i-x_j \parallel^2), \gamma>0 \quad \text{(Eq 9)}$$

Here x_i and x_j denote the *i*th and *j*th data point, respectively, and x_i^T is the transposed form of x_i.

The evaluation results are listed in Table 2. The RBF kernel function shows balanced and consistent predictive power, outperforming the two other kernel functions in both specificity and accuracy, while trailing only slightly behind the linear kernel function in sensitivity. Hence, subsequent SVM models reported in this study all employed the RBF kernel function.

Table 2. 10-fold cross-validation results of SVM modelling for kernel function selection.

	Sensitivity (%)	Specificity (%)	Accuracy (%)
Linear	94.60±1.44	31.11±4.63	73.71±0.68
Polynomial	56.93±7.35	91.07±2.66	68.16±4.36
RBF	81.87±3.66	93.02±11.26	85.54±5.58

Descriptor selection and comparison of classification algorithms

In this work, each protein sequence was represented as a 175-dimension descriptor vector (Materials and methods). As previously reported, choosing a relevant and complementary combination of descriptors for a model typically leads to better performance in machine learning approaches[14, 38], possibly resulting from the removal of noisy descriptors interfering with parameter optimization during model construction. To find the best combination of descriptors, two selection methods were implemented in this study, randomized and *F*-score based[31] (Materials and methods).

The two selection methods were implemented and evaluated according to Modeling Strategy II (Materials and methods) for three classification algorithms, SVM, NN and DT. The randomized descriptor selection method consistently outperformed the *F*-score based descriptor selection method in all 3 classification algorithms (Figure 2, Table S1). Consistent with our findings, other groups have also reported superior performance for the randomized descriptor selection method in machine-learning algorithms[14, 17]. Therefore, the randomized descriptor selection method was employed in all subsequent studies.

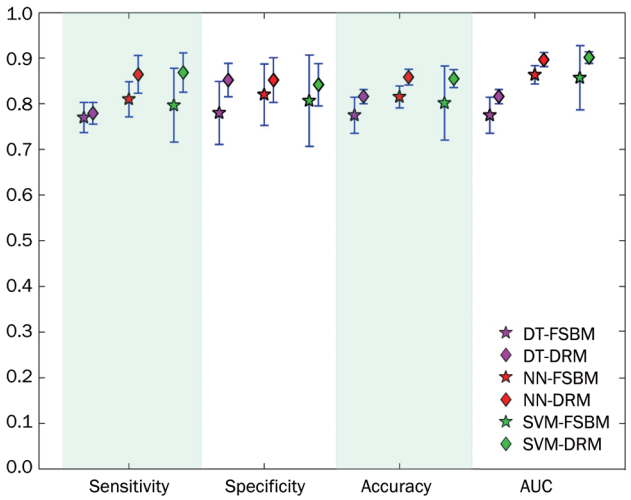

Figure 2. Comparison of three algorithms using two descriptor selection methods. FSBM, F-score based modeling; DRM, descriptor randomization modeling.

SVM and NN achieved similar performance and outperformed DT in all metrics except specificity (Figure 2). To further discriminate between SVM and NN, the extensibility of the model is evaluated, which is defined as the performance of other models constructed from the same training set but using other algorithms. Accordingly, 21 SVM models and 36 NN models were selected based on the criteria of Accuracy >0.80 and AUC>0.85. The training sets of the 21 SVM models were used to train 21 NN models and 21 DT models, and the performance metrics are illustrated in Figure 3A (Table S2). Likewise, the training sets for the 36 NN models were utilized to train 36 SVM models and 36 DT models (Figure 3B, Table S3). Subsequently, an ANOVA statistical test was implemented to analyze the difference between the performances of the models (Figure 4, Table S4), which demonstrates that the training sets for the 21 SVM models have better extensibility. The relevant ROC curves for the 21 SVM models are shown in Figure 5 (Table S5), and the AUCs of the 21 SVM models vary between 0.90 and 0.95, suggesting that each SVM model performs well. Therefore, we selected the 21 best-performing SVM models and used their corresponding training sets to train 21 NN models and 21 DT models. The performance metrics for the 63 models are shown in Table 3. For convenience, when the three models (one SVM model, one NN model and one DT model) are based on an identical training set, they are called "parallel models". The three algorithms and 21 parallel models were applied in subsequent studies.

Qualitative evaluation using multiple datasets

Three datasets were used to further assess our multi-algorithm and multi-model based strategy.

Phase II targets

Targets of drugs undergoing phase II clinical trials were collected and those that were included in the T-Set were

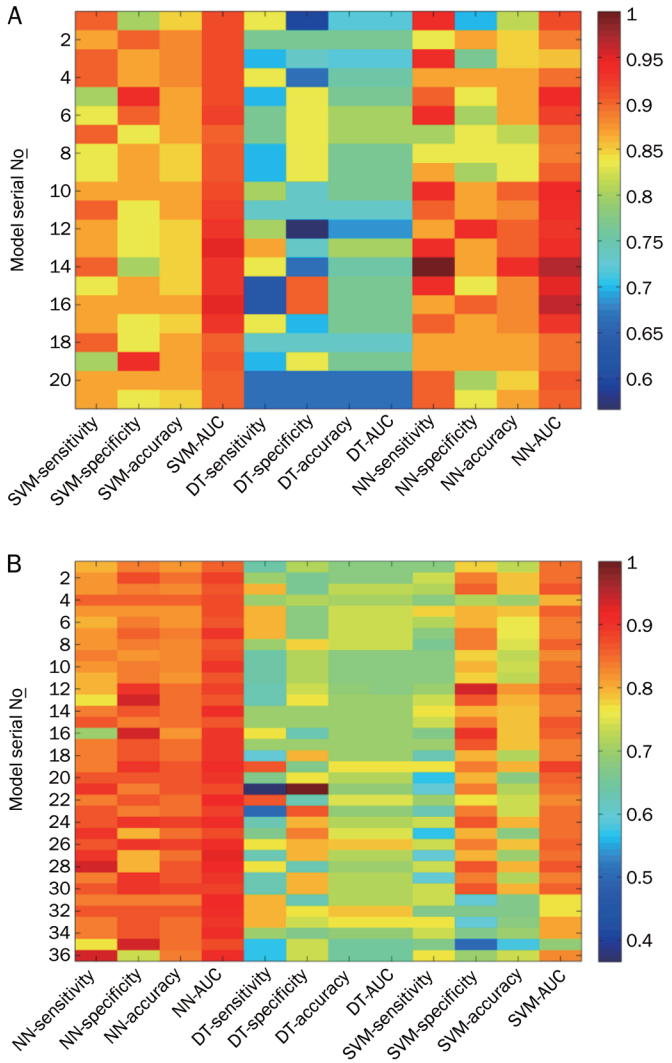

Figure 3. Evaluation of extensibility of the training sets of the 21 SVM models and the 36 NN models. The X-axis represents all of the performance metrics for the three algorithms, and the Y-axis is the model serial number. (A) Evaluation based on the training and testing sets of the 21 SVM models for the three algorithms. (B) Evaluation based on the training and testing sets of the 36 NN models for the three algorithms.

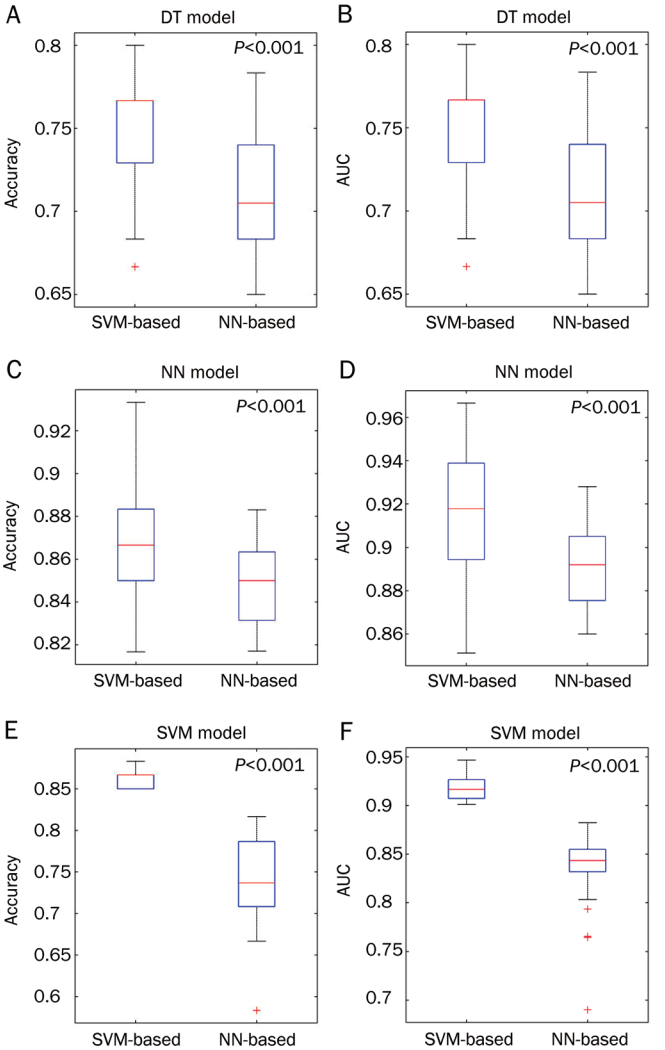

Figure 4. ANOVA statistical test. Analysis of differences in (A) the accuracies and (B) the AUCs between the DT models based on the training sets of the 21 SVM models and those of the 36 NN models. Analysis of differences in (C) the accuracies and (D) the AUCs between the NN models based on the training sets of the 21 SVM models and those of the 36 NN models. Analysis of differences in (E) the accuracies and (F) the AUCs between the SVM models based on the training sets of the 21 SVM models and those of the 36 NN models.

removed, resulting in 202 potential drug targets. As shown in Figures 6A and 6D (Table S6), all of the SVM and NN models produced consistent classifications, whereas the DT models, especially models 15 and 16, exhibited more variation. Nevertheless, the majority of the models classified over 40% of the clinical targets as true drug targets and nearly 60% as non-targets. Reports indicate that 66% of compounds entering phase II clinical trials fail prior to phase III[39] and that 30% of attritions in clinical trials are caused by a lack of efficacy[40], which can often be attributed to inappropriate targets. This finding qualitatively supports our results.

Human proteome

The whole human proteome dataset, including 20 331 proteins, was downloaded from Swiss-Prot[23]. After 306 targets originally included in T-Set were removed, all 63 models were applied to the remaining 20 025 proteins (Figures 6B and 6E). Most models predicted that at least 30% of proteins in the human proteome are drug targets, in qualitative agreement with other studies[3–5]. Again, DT-models 15 and 16 predicted a lower percentage of targets than other models and algorithms, suggesting that these two models should be utilized after more careful consideration. More detailed analyses and classification of the whole human proteome are provided later in this article.

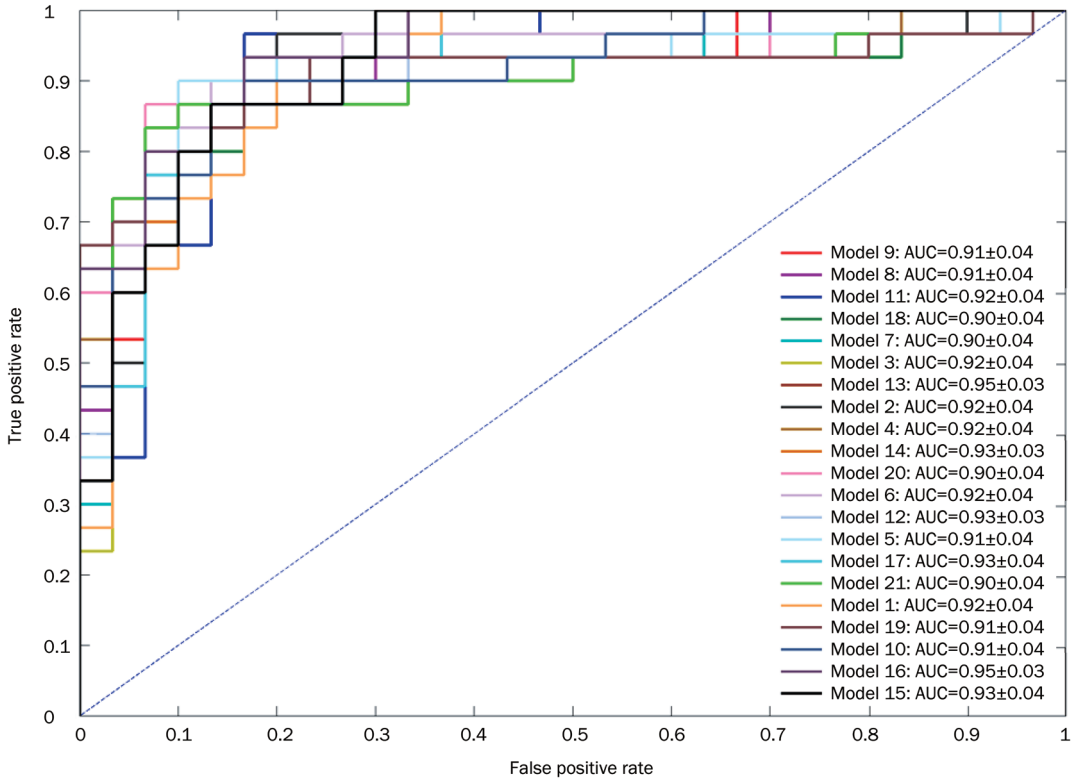

Figure 5. Receiver operating characteristic curves (ROCs) of the 21 SVM models.

Table 3. Performance metrics of 21 sorted parallel models for each of the three algorithms according to their ASE bars.

#Rank	ASE bar height	Model serial	#Descriptor	SVM				DT				NN			
				Sen	Spe	Acc	AUC	Sen	Spe	Acc	AUC	Sen	Spe	Acc	AUC
1	0.079	9	170	0.83	0.87	0.85	0.91	0.70	0.83	0.77	0.77	0.87	0.80	0.83	0.90
2	0.085	8	165	0.83	0.87	0.85	0.91	0.70	0.83	0.77	0.77	0.83	0.83	0.83	0.89
3	0.111	11	105	0.90	0.83	0.87	0.92	0.73	0.73	0.73	0.73	0.90	0.87	0.88	0.93
4	0.111	18	115	0.90	0.83	0.87	0.90	0.73	0.73	0.73	0.73	0.87	0.87	0.87	0.89
5	0.117	7	150	0.90	0.83	0.87	0.90	0.77	0.83	0.80	0.80	0.80	0.83	0.82	0.89
6	0.131	3	110	0.90	0.87	0.88	0.92	0.70	0.73	0.72	0.72	0.93	0.77	0.85	0.85
7	0.133	13	115	0.87	0.83	0.85	0.95	0.87	0.73	0.80	0.80	0.93	0.87	0.90	0.94
8	0.134	2	100	0.87	0.90	0.88	0.92	0.77	0.77	0.77	0.77	0.83	0.87	0.85	0.89
9	0.140	4	120	0.90	0.87	0.88	0.92	0.83	0.67	0.75	0.75	0.87	0.87	0.87	0.90
10	0.157	14	120	0.90	0.80	0.85	0.93	0.83	0.67	0.75	0.75	1.00	0.87	0.93	0.97
11	0.158	20	140	0.87	0.87	0.87	0.90	0.67	0.67	0.67	0.67	0.90	0.80	0.85	0.91
12	0.161	6	135	0.83	0.90	0.87	0.92	0.77	0.83	0.80	0.80	0.93	0.80	0.87	0.92
13	0.163	12	110	0.87	0.83	0.85	0.93	0.80	0.57	0.68	0.68	0.87	0.93	0.90	0.93
14	0.176	5	125	0.80	0.93	0.87	0.91	0.70	0.83	0.77	0.77	0.90	0.83	0.87	0.94
15	0.177	17	160	0.87	0.83	0.85	0.93	0.83	0.70	0.77	0.77	0.90	0.87	0.88	0.93
16	0.180	21	160	0.87	0.83	0.85	0.90	0.67	0.67	0.67	0.67	0.90	0.83	0.87	0.90
17	0.193	1	100	0.90	0.80	0.85	0.92	0.83	0.60	0.72	0.72	0.93	0.70	0.82	0.92
18	0.198	19	135	0.80	0.93	0.87	0.91	0.70	0.83	0.77	0.77	0.87	0.87	0.87	0.89
19	0.222	10	100	0.87	0.87	0.87	0.91	0.80	0.73	0.77	0.77	0.93	0.87	0.90	0.94
20	0.411	16	155	0.87	0.87	0.87	0.95	0.63	0.90	0.77	0.77	0.87	0.90	0.88	0.96
21	0.490	15	125	0.83	0.87	0.85	0.93	0.63	0.90	0.77	0.77	0.93	0.83	0.88	0.95

Sen, sensitivity; Spe, specificity; Acc, accuracy; AUC, area under the ROC curve.

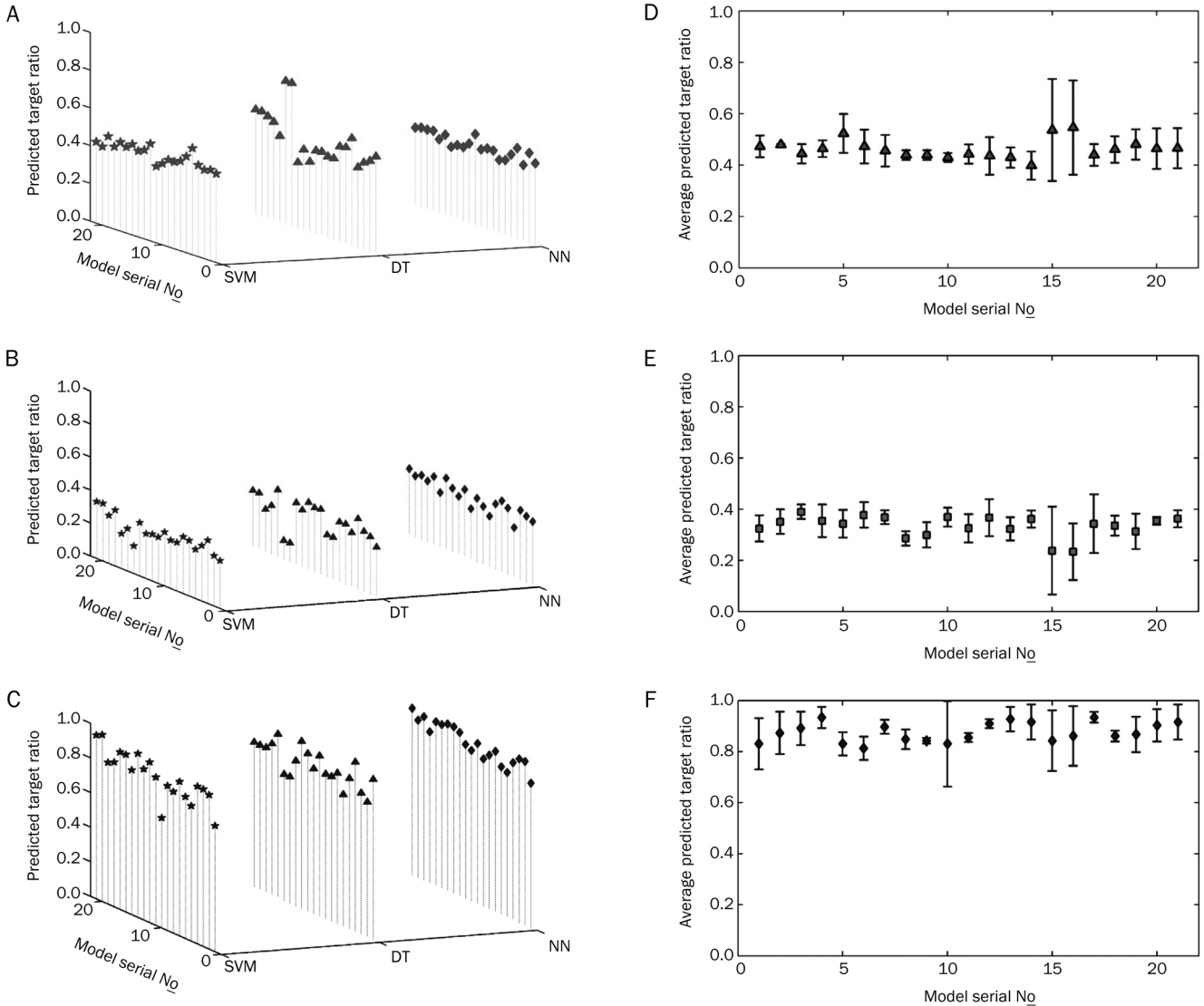

Figure 6. Evaluation of the 21 parallel models against three testing datasets. Evaluation against (A) Dataset I, clinical phase II targets (size: 202), (B) Dataset II, human proteome (size: 20 025), and (C) Dataset III, targets of withdrawn drugs (size: 55). Mean values and standard errors of the 21 models using the 3 algorithms against (D) Dataset I, (E) Dataset II, and (F) Dataset III.

Targets of withdrawn drugs in DrugBank

Among the 109 targets of withdrawn drugs obtained from DrugBank[4], 54 entries overlapped with the T-Set and were removed. All 63 models were applied to the remaining 55 targets. As shown in Figures 6C and 6F, models 10, 15, and 16 showed more variation, while other models produced more consistent results. The majority of our models predicted that ~85%–95% of targets of withdrawn drugs were true targets, suggesting that most of the withdrawals of marketed drugs may not have been caused by target druggability. This finding is intuitive because most marketed drugs should have demonstrated at least some efficacy to receive regulatory approval, which suggests the validity of the targets.

Quantitative evaluation with accumulated standard error

The above tests qualitatively demonstrated the consistency of the models and of the three algorithms. The Accumulated Standard Error (ASE) was used to provide a quantitative evaluation for the above tests (Equation 6, see Materials and methods).

The ASE bars are shown in Figure 7 (Table S7). The lower the ASE bar, the more robust the model. Models 8 and 9 exhibit the least discrepancy across the three algorithms, indicating that they are the most self-consistent models. On the contrary, models 15 and 16 are the most variable across the three algorithms. Twenty one models are ranked according to their ASE bars, and the ranked performance metrics of the

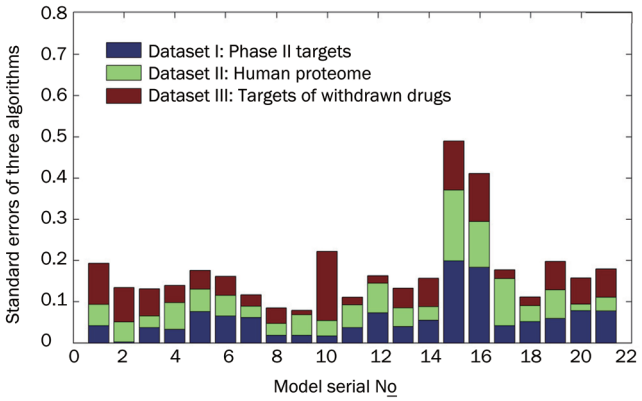

Figure 7. Bar chart of accumulated standard errors (ASE).

three algorithms are illustrated in Table 3. The subsequent applications of these models in the multi-algorithm and multi-model based strategy are based on their ranked order.

Assessment of the multi-algorithm and multi-model strategy

Next, we evaluated whether multi-algorithm and/or multi-model based strategies outperform single-algorithm and single-model based strategies. A graphical illustration of multi-algorithm and/or multi-model based strategies is given in Figure 8 (see Materials and methods for details). A total of 67 targets and 33 non-targets were selected randomly from the T-Set and NT-Set, respectively, and each of the strategies was applied to the combined set of 100 entries. This exercise was repeated 10 times, and each time a new test set was randomly selected. In the cases of single-algorithm and single-model based strategies (Figures 9A, 9B, and 9C, Table S8), the accuracy of most SVM and NN models was approximately 80%, but the error bars were relatively large. For comparison, the accuracy of almost all models utilizing a multi-algorithm based strategy and/or a multi-model based strategy was better than 80% (Figures 9D, 9E, and 9F). Furthermore, when multi-algorithm and multi-model based strategies were combined, the accuracy of target prediction increased to approximately 83%–85%, with higher consistency across the algorithms (Figure 9F). For instance, using 3 algorithms and the top 19 parallel models for each algorithm, the accuracy was over 85%. Therefore, the multi-algorithm and multi-model based strategy seems to be most reliable.

Based on simple sequence properties, Huang *et al*[22] and Li *et al*[17] also constructed SVM models for drug target prediction. When Huang *et al* applied the SVM method to predict the potential drug targets among ion channel proteins; the accuracy for a random dataset was ~50%. Even after optimization of description selection, the accuracy never increased beyond 80% for other datasets. Likewise, the accuracy of the SVM models developed by Li *et al* was less than 85% for both a carefully prepared testing dataset and a random dataset. The multi-algorithm and multi-model strategy has higher accuracy.

Evaluation of the multi-algorithm and multi-model strategy

Three separate datasets (Phase II, Phase III, and Phase IV) were prepared for further validation of the multi-algorithm and multi-model strategy. The proteins included in the Phase III and Phase IV datasets that overlapped with the Phase II dataset were removed from the Phase II dataset, and the pro-

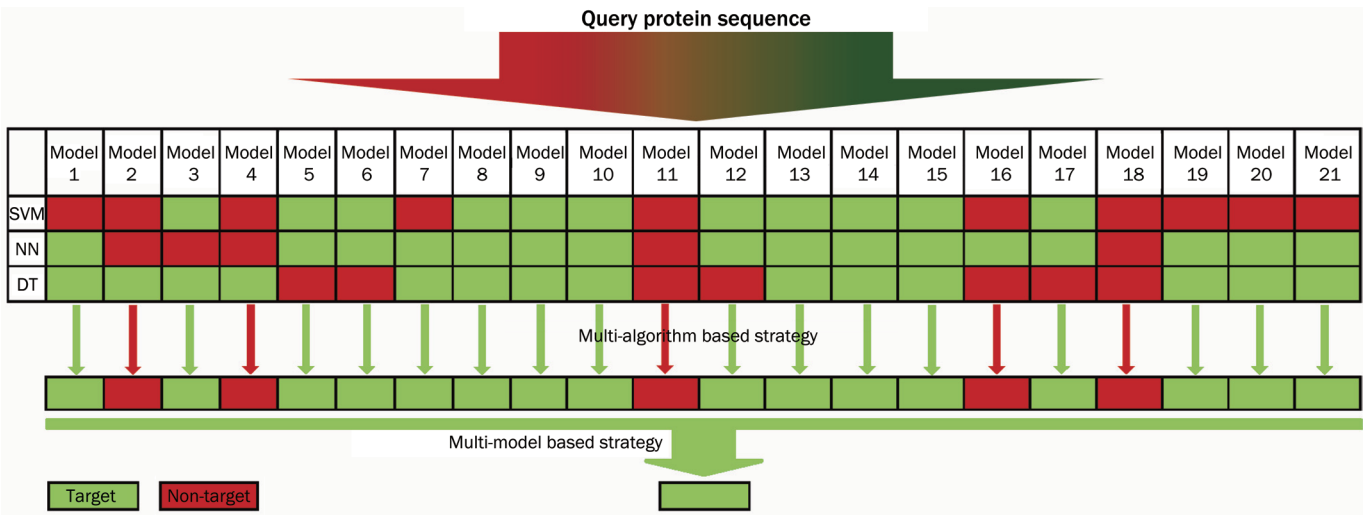

Figure 8. Illustration of multi-algorithm and/or multi-model based strategy. The red colored block represents a predicted non-target; the green colored block stands for a predicted target. Multi-algorithm based strategy: for *i* (*i*=1, 2, ⋯, 21), there are three corresponding models: SVM-model-*i*, NN-model-*i*, and DT-model-*i*. If a sequence is predicted as a target by no less than 2 models in the three models, the sequence is defined as a potential target. Multi-model based strategy: for algorithm *j* (*j*=SVM, NN, DT), there are *N* models (*N*=1, 2, ⋯, 21). If a sequence is predicted as a target by no less than [(*N*+1)/2] models, the sequence is defined as a potential target. Multi-algorithm and multi-model based strategy: successive combination of multi-algorithm based strategy and multi-model based strategy.

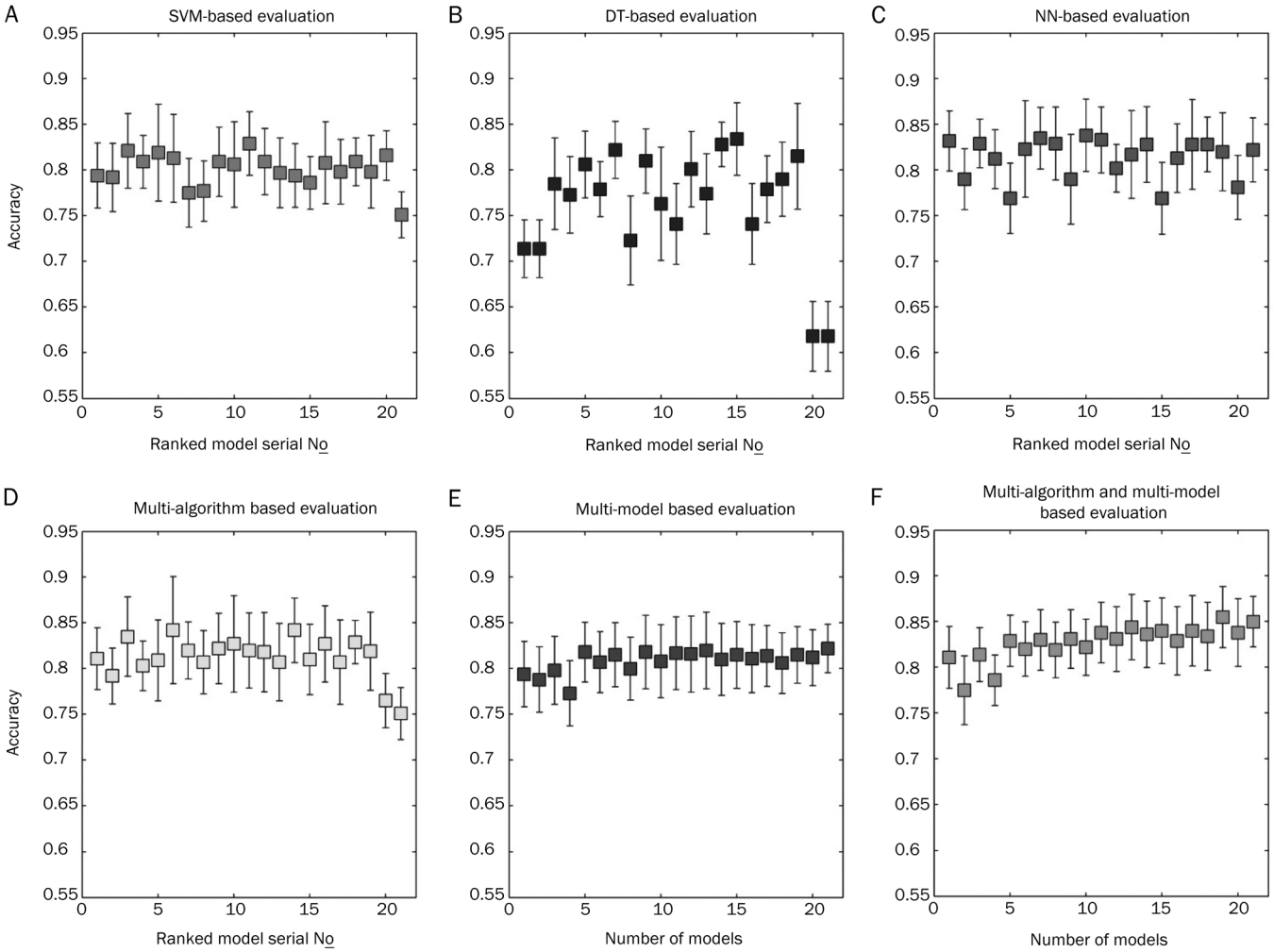

Figure 9. Multi-algorithm and/or multi-model based evaluation. Single-algorithm and single-model based evaluation using (A) the SVM algorithm, (B) the DT algorithm, and (C) the NN algorithm. (D) Multi-algorithm based evaluation. (E) Multi-model based evaluation. (F) Multi-algorithm and multi-model based evaluation.

teins included in the Phase IV dataset that overlapped with the Phase III data set were removed from the Phase III dataset. In addition, the sequences including unknown or nonstandard amino acids were removed from the three datasets. The rates of target identification in the three datasets were predicted with the 3 algorithms and the top 19 parallel models (Table 4). The target identification rates in the three datasets increased in the order Phase IV > Phase III > Phase II, which is consistent with the logical flow of R&D productivity in the pharmaceutical industry, indirectly supporting the practical utility of our approach.

Table 4. Evaluation of the multi-algorithm and multi-model strategy.

Dataset	Size	#Nontarget	#Quasi target	#Full target
Phase II	202	116/57.42%	51/25.25%	35/17.33%
Phase III	123	41/33.33%	47/38.21%	35/28.46%
Phase IV	181	41/22.65%	55/30.39%	85/46.96%

Novel drug target prediction

The potential drug targets in the human proteome were predicted with the 3 algorithms and the top 19 parallel models for each algorithm (multi-algorithm and multi-model based strategy). Any protein predicted to be a drug target by all 3*19 models was classified as a full target, which indicates a high confidence level; any protein validated as a potential drug target but that did not meet the criterion for a full target was classified as a quasi target. As shown in Table 5, 1932 (9.6%) of 20 025 human proteome proteins (excluding the T-Set targets) were predicted as full targets and 3990 (20.0%) were quasi targets (Dataset S1, Supporting information), suggesting that 29.6% of the proteins in the human proteome could be potential drug targets.

To analyze the distribution of target categories and to compare the distribution of all predicted targets and true targets, the true targets in the T-Set and all predicted targets were classified into 3 main categories and their corresponding subcategories based on the annotations of the UNIPROT[27] and

Table 5. Classification of the targets in the target dataset (T-Set) and the predicted targets in the human proteome*.

Category	Target dataset[a]	Full targets	Quasi targets	Potential targets[b]
Enzyme	118 (43.7%)	305 (19.5%)	1102 (66.6%)	1407 (43.7%)
Oxidoreductase	32 (11.9%)	47 (3.0%)	157 (9.5%)	204 (6.3%)
Transferase	39 (14.4%)	127 (8.1%)	416 (25.1%)	543 (16.9%)
Hydrolase	37 (13.7%)	109 (7.0%)	435 (26.3%)	544 (16.9%)
Lyase	4 (1.5%)	13 (0.8%)	28 (1.7%)	41 (1.3%)
Isomerase	6 (2.2%)	4 (0.3%)	10 (0.6%)	14 (0.4%)
Ligase	1 (0.4%)	10 (0.6%)	82 (5.0%)	92 (2.9%)
Receptor	141 (52.2%)	817 (52.3%)	297 (17.9%)	1114 (34.6%)
GPCR	83 (30.7%)	715 (45.8%)	27 (1.6%)	742 (23.1%)
Transporter[c]	48 (17.8%)	538 (34.4%)	355 (21.5%)	893 (27.8%)
Ionic channel[d]	28 (10.4%)	154 (9.9%)	84 (5.1%)	238 (7.4%)
Calcium channel	6 (2.2%)	31 (2.0%)	16 (1.0%)	47 (1.5%)
Chloride channel	5 (1.9%)	36 (2.3%)	10 (0.6%)	46 (1.4%)
Potassium channel	3 (1.1%)	29 (1.9%)	29 (1.8%)	58 (1.8%)
Sodium channel	4 (1.5%)	11 (0.7%)	4 (0.2%)	15 (0.5%)
Ligand-gated ion channel	13 (4.8%)	43 (2.8%)	19 (1.1%)	62 (1.9%)
Voltage-gated ion channel	10 (3.7%)	54 (3.5%)	42 (2.5%)	96 (3.0%)
Classified targets	270	1562	1655	3217
Unclassified targets[e]	36	370	2335	2705
Targets	306	1932	3990	5922
Total	306	20025	20025	20025

* Some proteins can be classified as at least two categories, so sum of all the classified categories is more than the number of classified targets. The percentage in the bracket is obtained from number of the classified targets divided by the corresponding number of the category, for example, 118 (43.7%) in the Target Dataset column is obtained from 118/270.

[a] Target Dataset (T-Set) composed of 306 targets (Figure 1B).

[b] Potential Targets are the sum of Full targets and Quasi targets.

[c] Transporter stands for any protein which are involved in importing, exporting or symporting any kinds of ions, sugars, peptides or proteins, *etc*.

[d] Ionic channel stands for any protein which is part of a transmembrane protein complex that forms a channel across the lipid bilayer through which specific inorganic ions can diffuse down their electrochemical gradients.

[e] Unclassified targets are those undetermined proteins that are not classified as the above enzymes, receptors or transporters.

Pfam[41] databases (Table 5). Receptor is the largest category among the true drug targets, followed by enzyme, representing 52.2% and 43.7% of the true drug targets, respectively. Similarly, receptor is also the largest category among the predicted full targets, where 52.3%, or 817 proteins, are receptors (including 715 GPCRs) and 538 are transporters. However, only 141 receptors (including 83 GPCRs) and 48 transporters are found in the T-Set, indicating a large undeveloped potential target space. Even if the 422 olfactory and 28 taste receptors are excluded from the original 715 GPCRs[42], there are still 292 GPCRs in the predicted potential targets for drug development. Thus, GPCRs are still one of the most important groups of drug targets[43]. Therewith, membrane proteins, such as GPCRs, transporters and ionic channels, should be prioritized for target validation studies. In addition, enzymes should not be neglected given their significant proportion among true drug targets. We classified 1407 enzymes as potential drug targets, among which transferases and hydrolases dominate. It is noteworthy that protein kinases, which belong to transferases, have successfully been targeted by drugs for several decades, and it is likely that this trend will continue in the future. Currently, dozens of inhibitors are undergoing clinical trials against protein kinases and several drugs have been launched commercially[44–50], demonstrating that protein kinases are one of the most important groups of drug targets. Therefore, enzymes, especially protein kinases, should also be emphasized in the pipeline for target validation.

Web server

We have implemented this work as a web server named D^3TPredictor. The server requires only a protein sequence as input and classifies it as a full target, a quasi target or a non-target. Users can fully customize the combination of algorithms and models for the prediction. Approximately 1600 tests, submitted by 40 internal and over 160 external users, have been completed, demonstrating that the D^3TPredictor web server is functional and stable. The server is available free of charge at http://www.d3pharma.com/d3tpredictor. This tool should be of significant value and interest to pharmaceutical research.

Discussion

The discovery of novel drug targets is of great importance in drug development, but it is laborious and costly. Hence, a reliable computational approach for drug target prediction would be of significant value. In this study, we carefully prepared the drug target and non-target datasets with multiple standards and selected appropriate kernel functions and descriptor selection approaches, which provide predictive models with superior reliability and robustness. Based on high-quality datasets, multiple models in combination with three algorithms (SVM, NN, and DT) were constructed. This approach was then evaluated qualitatively and quantitatively using three testing datasets, which are consistent with previously reported studies. Notably, we showed that the appropriate combination of multiple algorithms and multiple models yields better performance than individual models. Accordingly, we selected the best combination of 3 algorithms and 19 parallel models to predict potential drug targets in the human proteome. Approximately 30% of proteins in the human proteome were predicted to be potential drug targets, of which 1932 proteins were of high confidence level. Furthermore, the enrichment of GPCRs and kinases in the predicted targets agrees with the distribution of experimentally validated drug targets. In this regard, we suggest that GPCRs and kinases should be prioritized in future target validation studies.

Finally, we implemented our multi-algorithm and multi-model based strategy as a public web server, D^3TPredictor. To the best of our knowledge, D^3TPredictor is the first public web server for drug target prediction using a multi-algorithm and multi-model strategy. In addition, D^3TPredictor has been tested online internally and externally, highlighting its function and stability. This server should facilitate new advances in pharmaceutical research.

Acknowledgements

This work was supported by National Natural Science Foundation of China (81273435 and 21021063), National Science & Technology Projects (2012ZX09301001-004, 2012AA01A305, and 2013ZX09103001-001). Computational resources were provided by supercomputer TianHe-I in Tianjin and the Shanghai Supercomputing Center (SCC). The authors thank the developers of free and/or open source software for academic use, including SignalP-3.0, netOglyc-3.1d, netNglyc-1.0, tmhmm-2.0c and EMBOSS-6.0.1.

Author contribution

Wei-liang ZHU and Ji-ye SHI conceived and designed the research; Ying-tao LIU, Yi LI, and Zi-fu HUANG performed the research; Ying-tao LIU, Yi LI, Zi-fu HUANG, Ji-ye SHI, and Wei-liang ZHU wrote the paper; Zhi-jian XU, Zhuo YANG, Zhu-xi CHEN, and Kai-xian CHEN participated in algorithm selection and model construction.

Supplementary information

Dataset S1. Classification of predicted targets from 20025 proteins in the human proteome using the multi-algorithm and multi-model strategy presented in this work (EXCEL).

Tables S1–S8. Quantitative results with tables from Figures 2, 3, 4, 5, 6, 7, and 9 (DOC).

Supplementary information is available at Acta Pharmacologica Sinica's website.

References

1 Ohlstein EH, Ruffolo RR, Elliott JD. Drug discovery in the next millennium. Annu Rev Pharmacol Toxicol 2000; 40: 177–91.
2 Hopkins AL, Groom CR. The druggable genome. Nat Rev Drug Discov 2002; 1: 727–30.
3 Drews J. Drug discovery: a historical perspective. Science 2000; 287: 1960–4.
4 Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, *et al*. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res 2006; 34: D668–72.
5 Drews J. Genomic sciences and the medicine of tomorrow. Nat Biotechnol 1996; 14: 1516–8.
6 Overington JP, Al-Lazikani B, Hopkins AL. Opinion — How many drug targets are there? Nat Rev Drug Discov 2006; 5: 993–6.
7 Butcher SP. Target discovery and validation in the post-genomic era. Neurochem Res 2003; 28: 367–71.
8 An J, Totrov M, Abagyan R. Comprehensive identification of "druggable" protein ligand binding sites. Genome Inform 2004; 15: 31–41.
9 Russ AP, Lampel S. The druggable genome: an update. Drug Discov Today 2005; 10: 1607–10.
10 Hardy LW, Peet NP. The multiple orthogonal tools approach to define molecular causation in the validation of druggable targets. Drug Discov Today 2004; 9: 117–26.
11 Hajduk PJ, Huth JR, Tse C. Predicting protein druggability. Drug Discov Today 2005; 10: 1675–82.
12 Hajduk PJ, Huth JR, Fesik SW. Druggability indices for protein targets derived from NMR-based screening data. J Med Chem 2005; 48: 2518–25.
13 Mullner S, Neumann T, Lottspeich F. Proteomics — a new way for drug target discovery. Arzneimittelforschung 1998; 48: 93–5.
14 Han LY, Zheng CJ, Xie B, Jia J, Ma XH, Zhu F, *et al*. Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness. Drug Discov Today 2007; 12: 304–13.
15 Bakheet TM, Doig AJ. Properties and identification of human protein drug targets. Bioinformatics 2009; 25: 451–7.
16 Xu H, Lin M, Wang W, Li Z, Huang J, Chen Y, *et al*. Learning the drug target-likeness of a protein. Proteomics 2007; 7: 4255–63.
17 Li Q, Lai L. Prediction of potential drug targets based on simple sequence properties. BMC Bioinformatics 2007; 8: 353.
18 Zhang GL, Khan AM, Srinivasan KN, August JT, Brusic V. Neural models for predicting viral vaccine targets. J Bioinform Comput Biol 2005; 3: 1207–25.
19 Niwa T. Prediction of biological targets using probabilistic neural networks and atom-type descriptors. J Med Chem 2004; 47: 2645–50.
20 Nidhi, Glick M, Davies JW, Jenkins JL. Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. J Chem Inf Model 2006; 46: 1124–33.
21 Xu H, Fang Y, Yao L, Chen Y, Chen X. Does drug-target have a likeness? Method Inf Med 2007; 46: 360–6.
22 Huang C, Zhang R, Chen Z, Jiang Y, Shang Z, Sun P, *et al*. Predict potential drug targets from the ion channel proteins based on SVM. J Theor Biol 2009; 262: 750–6.

23 Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003; 31: 365–70.
24 Plosker GR. Information strategist — Thomson pharma and infotrieve life science research center: New directions for online aggregators. Online 2006; 30: 47–51.
25 Ji ZL, Han LY, Yap CW, Sun LZ, Chen X, Chen YZ. Drug adverse reaction target database (DART): proteins related to adverse drug reactions. Drug Saf 2003; 26: 685–90.
26 Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, *et al*. The protein data bank. Nucleic Acids Res 2000; 28: 235–42.
27 Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, *et al*. The universal protein resource (UniProt). Nucleic Acids Res 2005; 33: D154–9.
28 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990; 215: 403–10.
29 Vanopdenbosch N, Cramer R, Giarrusso FF. Sybyl, the integrated molecular modeling system. J Mol Graph 1985; 3: 110–1.
30 Halgren TA. Identifying and characterizing binding sites and assessing druggability. J Chem Inf Model 2009; 49: 377–89.
31 Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM TIST 2011; 2: 1–27.
32 Rice P, Longden I, Bleasby A. EMBOSS: the european molecular biology open software suite. Trends Genet 2000; 16: 276–7.
33 Nielsen H, Engelbrecht J, Brunak S, vonHeijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng 1997; 10: 1–6.
34 Center for biological sequence analysis [homepage on the Internet]. Technical University of Denmark; c2001–2013 [updated 2013 Jun 5; cited 2013 Jul 19]. Available from: http://www.cbs.dtu.dk/services/NetNGlyc/.
35 Julenius K, Molgaard A, Gupta R, Brunak S. Prediction, conservation analysis, and structural characterization of mammalian mucin-type *O*-glycosylation sites. Glycobiology 2005; 15: 153–64.
36 Li ZR, Lin HH, Han LY, Jiang L, Chen X, Chen YZ. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res 2006; 34: W32–7.
37 Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982; 157: 105–32.
38 Dobson PD, Doig AJ. Distinguishing enzyme structures from non-enzymes without alignments. J Mol Biol 2003; 330: 771–83.
39 Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, *et al*. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov 2010; 9: 203–14.
40 Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov 2004; 3: 711–5.
41 Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL. The Pfam protein families database. Nucleic Acids Res 2000; 28: 263–6.
42 Lagerstrom MC, Schioth HB. Structural diversity of G protein—coupled receptors and significance for drug discovery. Nat Rev Drug Discov 2008; 7: 339–57.
43 Chantry D. G protein—coupled receptors: from ligand identification to drug targets. 14–16 October 2002, San Diego, CA, USA. Expert Opin Emerg Drugs 2003; 8: 273–6.
44 Cohen P. Protein kinases — the major drug targets of the twenty — first century? Nat Rev Drug Discov 2002; 1: 309–15.
45 Asano T, Ikegaki I, Satoh S, Seto M, Sasaki Y. A protein kinase inhibitor, fasudil (AT-877): A novel approach to signal transduction therapy. Cardiovasc Drug Rev 1998; 16: 76–87.
46 Garber K. Rapamycin's resurrection: a new way to target the cancer cell cycle. J Natl Cancer Inst 2001; 93: 1517–9.
47 Schindler T, Bornmann W, Pellicena P, Miller WT, Clarkson B, Kuriyan J. Structural mechanism for STI-571 inhibition of abelson tyrosine kinase. Science 2000; 289: 1938–42.
48 Sebolt-Leopold JS, Dudley DT, Herrera R, Van Becelaere K, Wiland A, Gowan RC, *et al*. Blockade of the MAP kinase pathway suppresses growth of colon tumors *in vivo*. Nat Med 1999; 5: 810–6.
49 Senderowicz AM. Small molecule modulators of cyclin-dependent kinases for cancer therapy. Oncogene 2000; 19: 6600–6.
50 Morin MJ. From oncogene to drug: development of small molecule tyrosine kinase inhibitors as anti-tumor and anti-angiogenic agents. Oncogene 2000; 19: 6574–83.