Support vector machines approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness

Lian Yi Han, Chan Juan Zheng, Bin Xie, Jia Jia, Xiao Hua Ma, Feng Zhu, Hong Huang Lin, Xin Chen and Yu Zong Chen

Bioinformatics and Drug Design Group, Department of Pharmacy and Department of Computational Science, National University of Singapore, Blk Soc 1, Level 7, 3 Science Drive 2, Singapore 117543

Identification and validation of viable targets is an important first step in drug discovery and new methods, and integrated approaches are continuously explored to improve the discovery rate and exploration of new drug targets. An *in silico* machine learning method, support vector machines, has been explored as a new method for predicting druggable proteins from amino acid sequence independent of sequence similarity, thereby facilitating the prediction of druggable proteins that exhibit no or low homology to known targets.

Introduction

Most drugs exert their therapeutic effect by binding to and regulating the activity of a particular protein or nucleic acid target. The identification and validation of such targets is an important first step in the drug discovery processes [1,2], and various target identification technologies [3,4] have been developed by analyzing disease relevance, functional roles, expression profiles and loss-of-function genetics between normal and disease states [5–12]. Computational methods have been used for predicting druggable proteins, the activity of which can be regulated by drug-like molecules [13], from genomic, structural and functional information [13–16] – druggable proteins with key roles in a disease can then be explored as therapeutic targets [13].

Despite increasing levels of spending and extensive use of new technologies [17], there is a shortage of approved drugs and a lack of innovative drug targets [5]. Therefore, new and improved methods [18], and integrated and systems-based approaches [5,6,19], are being explored for identifying targets and druggable proteins. The commonly used computational methods have primarily been based on the detection of sequence and functional similarity to known targets [13,14], drug-binding domain family affiliation [7,13], and structural analysis of geometric and energetic features [15,16]. These methods are less effective in finding targets that exhibit no or low homology to known targets, disease proteins and proteins with available 3D structures. However, such non-homologous and structurally unknown proteins constitute a substantial percentage, ~20–100%, of the open reading frames in many of the completed genomes and might, therefore, be an untapped source of novel drug targets [20]. Hence, methods independent of sequence and functional similarity, and structural availability, are highly desirable.

One such method, support vector machines (SVMs), which is summarized in Box 1, has recently been explored for predicting druggable proteins [21], anticancer genes [22], proteins in families of high target concentrations [23–28], as well as proteins of various broadly defined functional and structural classes [29], from sequence-derived constitutional and physicochemical properties, irrespective of similarity to known proteins. This method is particularly useful for predicting novel druggable proteins that exhibit no or low homology to known targets. Here, we describe SVM algorithms, evaluate their performance through statistical and proof-of-concept tests, and discuss the underlying difficulties and perspectives of their potential applications for facilitating the discovery of innovative targets.

Recent progress in exploring SVM approaches for predicting druggable proteins

To facilitate the identification of genes related to anticancer targeting genes, SVMs have been used for assigning genes into predefined

Corresponding author: Chen, Y.Z. (phacyz@nus.edu.sg)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2007.02.015

BOX 1

SVMs

SVMs represent a supervised ML method for classifying objects into separate groups and for developing regression models based on quantitative information on the characteristics inherent in a training set of objects. Regression models describe quantitative relationships between the activities and characteristics of objects. SVMs use the structural risk minimization principle to minimize both training and generalization errors. When used for classification, SVMs separate positive and negative training objects by projecting their characteristics into a multidimensional feature space and then constructing a hyperplane that separates these positive and negative samples optimally. A testing sample is then projected onto this multidimensional space to determine its class affiliation based on its relative position to the hyperplane.

SVMs have been explored for predicting druggable proteins, compounds that interact with specific target or absorption, distribution, metabolism, excretion and toxicity (ADMET)-related protein, compounds of specific pharmacokinetic or toxicological properties, and peptide vaccines. Other applications include prediction of proteins of specific structural or functional class, prediction of protein–protein interactions, protein fold recognition, prediction of a specific class of RNAs, disease diagnosis or outcome prediction, microarray gene expression data analysis, text categorization, hand-writing recognition, tone recognition, image classification and recognition, sonar data analysis, vehicle identification, and flood stage forecasting.

mechanistic categories based on their expression and anticancer profiles, enabling the selection of genes related to the common anticancer mechanisms, with a 2–10-fold lower false-positive rate against random selection [22]. Known drug targets are strongly represented in families such as G-protein-coupled receptors (GPCRs), nuclear receptors, DNA-binding proteins, specific transporter classes (e.g. ion transport proteins) and enzyme families (e.g. kinases) [21]. SVMs have consistently shown good performance for predicting members of these families [23–28]. A twofold cross-validation test suggests that SVMs recognize members of GPCR subfamilies at a lower per-sequence error rate than those of BLAST and the hidden Markov model [23]. SVMs have also performed well for proteins of less than 20% sequence identity [25]. These studies suggest that SVMs might be useful for finding novel druggable proteins.

Efforts have also been directed at predicting druggable proteins based on common features of known druggable targets instead of affiliation to a specific mechanism or family [18,21]. Investigations of known drug targets have shown that they have common characteristics collectively manifested by some combinations of functional, structural, physicochemical and localization features [13,19]. SVMs have consistently shown good performance in predicting various classes of proteins of specific functional, structural, physicochemical or localization feature [29], which is why one would expect them to be similarly useful for predicting druggable proteins characterized by combinations of these features. Recently reported proof-of-principle studies have shown that SVMs are capable of predicting druggable proteins at reasonably good accuracy levels [18,21].

SVM approach for predicting druggable proteins

Outline of prediction strategy

One strategy for predicting druggable proteins from their sequences, without the use of sequence similarity, is to use a sequence-independent classifier generated from the analysis of known druggable targets that share some characteristics but might be substantially different in sequence, structure and function [13,21]. To be of value for therapeutic intervention, targets need to have crucial roles in disease processes. It is preferable that they are not involved in other important physiological processes, to rule out side effects. Expression of these targets should be either tissue selective or at lower levels than that coverable by drugs at therapeutic dosage, to enable sufficient drug efficacy. To be druggable, targets need to contain drug-binding sites with certain structural and physicochemical properties to accommodate high-affinity, site-specific binding and subsequent activity modulation by drug-like molecules. These characteristics define the sequence, genomic, structural and proteomic profiles of druggable proteins and the roles of targets at the pathway, cellular and physiological levels.

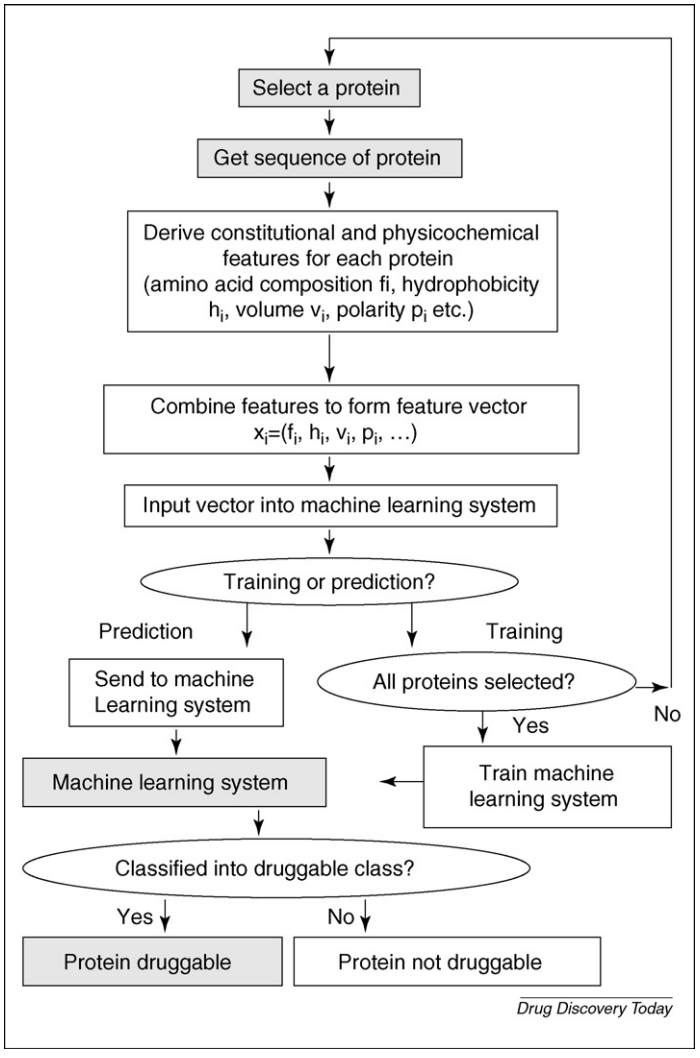

FIGURE 1

Schematic diagram illustrating the process of the prediction of a druggable protein from the sequence of a protein by using an ML method.

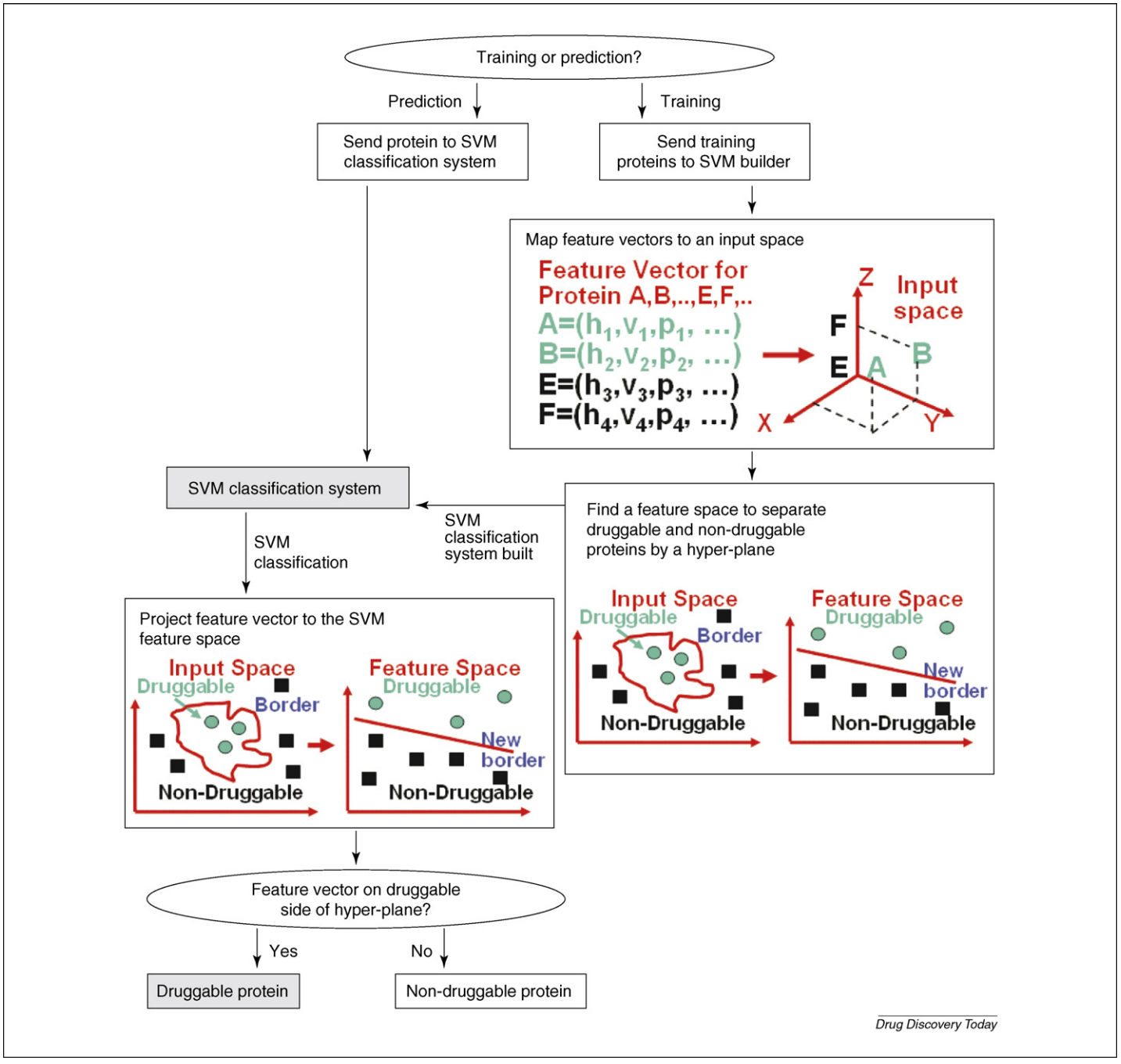

FIGURE 2

Schematic diagram illustrating the process of the prediction of a druggable protein from the sequence of a protein by using SVMs. A,B: feature vectors of druggable proteins; E,F: feature vectors of non-druggable proteins; green circles: druggable proteins; black-filled squares: non-druggable proteins; feature vector (h_j, p_j, vj,. . .) represents hydrophobicity, volume, polarizability, etc.

Proteins can be divided into druggable and non-druggable classes. Thus, two-tier classification machine learning (ML) methods such as SVMs [30] can be applied for developing an artificial intelligence model to separate druggable from non-druggable proteins. Figure 1 illustrates the process of training an ML model and using it for predicting druggable proteins. Each protein is represented by a feature vector composed of sequence-derived descriptors representing its structural and physicochemical properties. As illustrated in Figure 2, SVMs classify proteins by projecting their feature vectors into a multidimensional space in which druggable and non-druggable proteins are separated by a hyperplane. A protein is predicted to be druggable or non-druggable depending on whether its feature vector is projected onto the druggable or non-druggable side of the hyperplane.

Sources of druggable and non-druggable proteins

Sufficiently diverse sets of druggable and non-druggable proteins are needed for training and testing a SVM prediction model. There are 1484 successfully commercialized and research targets in the therapeutic target database [31] with available sequence informa-

tion, which together form the druggable class. Some viral and microbial targets have multiple sequence entries because there are substantial sequence variations across strains. Based on their family distribution pattern, targets are expected to be represented by <800 protein families, including 460 covered by the known targets [21,32]. There are 8957 protein families in the Protein Family (Pfam) database [33] that contain no known target at present. Protein families in the Pfam database are defined based on domain affiliations or sequence clustering. Therefore, without substantially reducing SVM prediction performance, putative non-druggable proteins can be tentatively derived from these non-target families, producing a maximum possible 'wrong' family representation rate of <7%, even when all of the <340 unidentified target families are misplaced [21]. Representative proteins of these non-target families form the non-druggable class. Importantly, inclusion of the representative of a 'wrong' family into the non-druggable class does not preclude other family members from being classified as druggable. Statistically, a substantial percentage of druggable members can be located on the druggable side of the SVM hyperplane, even if its family representative is on the non-druggable side. Therefore, in principle, a reasonably good SVM prediction model can be derived from these putative non-druggable proteins for predicting druggable proteins rather than Pfam family membership, as confirmed by the case studies described here. The quality of the non-druggable class and the performance of SVM can be further improved, in line with the discovery of new targets.

Representation of protein sequence

In using SVMs for predicting druggable proteins, each protein is represented by a multidimensional feature vector composed of descriptors that encode constitutional and physicochemical properties of that protein [29]. These descriptors define structural, functional and interactive properties of proteins. Web servers such as PROFEAT [34] and ProtParam [35] have been developed by different research groups for facilitating the computation of these descriptors. The descriptors used for predicting druggable proteins [21] include a constitutional descriptor, amino acid composition and several physicochemical descriptors that describe the composition, transition and distribution of hydrophobicity (*h*), polarity (*p*), polarizability (*z*), charge (*c*), secondary structures (*s*), solvent accessibility (*a*), surface tension (*t*) and normalized van der Waals volumes (*v*) [36].

Amino acid composition is the fraction of each type of amino acid in a sequence $f_k = N_k/N$, where $k = 1,2,3, \ldots, 20$ is the index of amino acids, N_k is the number of a particular type of amino acid and N is sequence length. For computing descriptors of each of the physicochemical properties of a protein, amino acids are divided into three types. For instance, for hydrophobicity descriptors, amino acids can be divided into hydrophobic (CVLIMFW), neutral (GASTPHY) and polar (RKEDQN) types. Three descriptors, composition (C_q), transition (T_q) and distribution (D_q), are introduced to describe global composition of each of the physicochemical properties, where $q = h, p, z, c, s, a, t$ and v.

$$C_q = \left(\frac{N_{q_1}}{N}, \frac{N_{q_2}}{N}, \frac{N_{q_3}}{N}\right) \qquad \text{[Equation 1]}$$

Equation 1 represents the percentage of each type of residue in a sequence, where *Nqi* is the number of type i residues, and Equation 2 characterizes the percentage frequency of transition between different types of residues.

$$T_q = \left(\frac{T_{q_{12}}}{N-1}, \frac{T_{q_{13}}}{N-1}, \frac{T_q}{N-1}\right) \qquad \text{[Equation 2]}$$

Tq_{ij} is the number of type i to j transitions and *N-1* is the total number of transitions.

$$T_{qij} = T_{qij} \text{ and } D_q = (D_{q_1}, D_q, D_q) \qquad \text{[Equation 3]}$$

$$\text{with } D_{qi} = \left(\frac{P_{q_{i0}}}{N}, \frac{P_{q_{i25}}}{N}, \frac{P_{q_{i50}}}{N}, \frac{P_{q_{i75}}}{N}, \frac{P_{q_{i100}}}{N}\right)$$

These transitions are undirected, such that Equation 3 measures the chain length, within which the first 25%, 50%, 75% and 100% of the amino acids of a particular group are located, respectively, and where P_{qik} is the length, within which k% of type i residues are located. Overall, each physicochemical property is represented by 21 elements: three for C_q, three for T_q and 15 for D_q. The complete feature vector consists of 188 elements, including 20 for amino acid composition and 8×21 for physicochemical properties. All generated vectors have equal length.

SVM algorithms and software

SVMs represent a supervised ML method based on the structural risk minimization principle for minimizing both training and generalization error [30]. As illustrated in Figure 2, when used for classification, SVMs separate positive (druggable) and negative (non-druggable) training samples in a multidimensional space by constructing a hyperplane optimally positioned between the positive and negative samples. A testing sample is then projected onto this multidimensional space to determine its class affiliation based on its relative position to the hyperplane (druggable if on the druggable side of the hyperplane, non-druggable if on the non-druggable side of the hyperplane).

There are linear and non-linear SVMs: linear SVMs are applicable for samples separable by linear mapping of their feature vectors; non-linear SVMs are used for samples inseparable by linear mapping of their feature vectors, which is more useful for classifying proteins of diverse sequences and has been used primarily for predicting druggable [21] and other classes [29] of proteins. In non-linear SVMs, each feature vector x_i is projected into a higher dimensional feature space by using a kernel function such as a Gaussian function (Equation 4)

$$K(x_i, x_j) = e^{-||x_j - x_i||^2/2\sigma^2} \qquad \text{[Equation 4]}$$

where a hyperplane is constructed by finding a vector w and a parameter b minimize $||W||^2$. The hyperplane equation is shown in Equation 5

$$w^{\bullet} x_i + b = 0 \qquad \text{[Equation 5]}$$

and the equations for druggable (Equation 6) or non-druggable (Equation 7) are:

$$w^{\bullet} x_i + b \geq +1, \text{for } y_i = +1 \qquad \text{[Equation 6]}$$

$$w^{\bullet} x_i + b \leq -1, \text{for } y_i = -1 \qquad \text{[Equation 7]}$$

TABLE 1

Performance of ML methods SVM, PNN, kNN and decision tree, and sequence alignment method BLAST for predicting druggable proteins, as tested by a fivefold cross-validation study of 1484 druggable and 6637 non-druggable proteins. The number of druggable and non-druggable proteins in each training and testing set is described in the text

Method	Accuracy for druggable proteins P+ (%)						Accuracy for non-druggable proteins P– (%)						Average accuracy P (%)					
	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Average	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Average	Fold-1	Fold-2	Fold-3	Fold-4	Fold-5	Average
SVM	69.2	71.0	64.1	66.6	67.1	67.6	85.0	85.1	85.8	85.1	85.3	85.3	83.5	83.8	83.7	83.2	83.7	83.6
PNN	64.5	64.5	58.5	61.1	63.5	62.4	81.4	81.6	81.5	83.5	81.3	81.9	79.8	80.0	79.3	81.3	79.7	80.0
kNN	67.6	69.3	63.5	64.6	68.6	66.7	75.9	76.2	76.6	78.4	77.2	76.9	75.1	75.6	75.4	77.0	76.4	75.9
Decision tree	57.5	56.3	52.5	53.5	55.2	55.0	81.8	82.1	82.2	82.1	83.3	82.3	79.4	79.7	79.3	79.2	80.8	79.7
BLAST	61.9	63.5	60.8	65.0	61.0	62.4	99.9	100.0	99.9	99.9	99.8	99.9	96.2	96.6	96.1	96.4	96.3	96.3

Based on the derived w and b, a new protein, x, can be classified as druggable or non-druggable when the sign of $[(w^{\bullet}x) + b]$ is positive or negative. In the case studies described here, Gaussian kernel SVMs implemented by our own software [20] were used and were optimized by scanning the parameter σ. Free SVM software tools such as SVMlight, LIBSVM and mySVM can also be used. Websites for these and other free ML software tools are provided in the supplementary material. Because of the sequence diversity of druggable proteins, convergence parameters of these software tools might need to be adjusted to achieve a better prediction performance.

Performance measurement

The performance of SVMs has frequently been measured by the fivefold cross-validation method [29]. In this method, proteins in both the druggable and non-druggable classes are randomly divided into five subsets of approximately equal size. Four subsets are selected as the training set for developing SVM model, and the fifth as the testing set for evaluating it. This process is repeated five times so that every subset is used as a testing set once. The performance of SVM can be measured by positive accuracy (druggable; Equation 8), negative accuracy (non-druggable; Equation 9) and and overall accuracy (Equation 10).

$$P_{+} = TP/(TP + FN) \qquad \text{[Equation 8]}$$

$$P_{-} = TN/(TN + FP) \qquad \text{[Equation 9]}$$

$$P = (TP + TN)/N \qquad \text{[Equation 10]}$$

Here, *TP, TN, FP, FN* and *N* are the true positive, true negative, false positive, false negative and the total number of proteins, respectively [29].

Case studies for testing SVM prediction of druggable proteins

Several case studies were conducted for testing the ability of SVMs to predict druggable proteins. These include statistical estimation of prediction performance, comparison with the reported prediction results of other methods, and practical usefulness for genome searching of druggable proteins.

Statistical tests

Table 1 gives the performance of SVM prediction of druggable proteins based on a fivefold cross-validation study of 1484 druggable and 6637 non-druggable proteins. The five training sets consist of 1187/5310, 1187/5311, 1187/5311, 1187/5311 and 1188/5312 druggable/non-druggable proteins; and the corresponding testing sets contain 297/1328, 297/1327, 297/1327, 297/1327 and 296/1326 druggable/non-druggable proteins, respectively. These testing sets include 82, 72, 71, 74 and 64 'novel' druggable proteins, with none of their Pfam family members in the corresponding training set. The computed prediction accuracies for druggable and non-druggable proteins are in the range 64.1–71.0% and 85.0–85.8%, respectively. In particular, 44%, 50%, 43%, 32% and 39% of the 'novel' druggable proteins are correctly predicted, suggesting that SVMs are capable of predicting druggability beyond protein family membership. The prediction

TABLE 2

Known, successfully commercialized research and proposed research targets in *M. tuberculosis* genome

Gene name	Swissprot protein accession number	Target status [40]	SVM prediction result	Prediction result from DDA analysis [40]
rpoB	P0A680	Successfully commercialized	Druggable	Non-druggable
embC	P72059	Successfully commercialized	Druggable	Non-druggable
embA	P0A560	Successful target	Druggable	Non-druggable
embB	P72030	Successfully commercialized	Druggable	Non-druggable
rpsL	P41196	Successful target	Druggable	Non-druggable
inhA	P0A5Y6	Successful target	Druggable	Druggable
gyrA	Q07702	Successful target	Druggable	Druggable
gyrB	P41514	Successful target	Druggable	Druggable
alr	P0A4X2	Successful target	Druggable	Druggable
ddlA	P95114	Successfully commercialized	Druggable	Non-druggable
dfrA	P0A546	Research target	Druggable	Druggable
folP1	P0A578	Research target	Druggable	Druggable
fbpC	P0A4V4	Research target	Druggable	Non-druggable
fbpB	P31952	Research target	Druggable	Non-druggable
fbpD	P0A4V6	Research target	Non-druggable	Druggable
fbpA	P0A4V2	Research target	Druggable	Non-druggable
cyp51	P0A512	Research target	Non-druggable	Druggable
cyp121	P0A514	Research target	Druggable	Druggable
def	P96275	Research target	Druggable	Non-druggable
atpE	P63691	Research target	Non-druggable	Non-druggable
icl	P0A5H3	Proposed research target	Druggable	Non-druggable
pcaA	Q7D9R5	Proposed research target	Non-druggable	Non-druggable
relA	P66014	Proposed research target	Non-druggable	Non-druggable
devR	P95193	Proposed research target	Non-druggable	Non-druggable
devS	P95194	Proposed research target	Non-druggable	Non-druggable
lysA	P0A5M4	Proposed research target	Non-druggable	Druggable
panD	P65660	Proposed research target	Non-druggable	Druggable
panC	P0A5R0	Proposed research target	Non-druggable	Non-druggable
glnE	P69942	Proposed research target	Non-druggable	Non-druggable
glnA1	P0A590	Proposed research target	Druggable	Non-druggable
aroK	P0A4Z2	Proposed research target	Non-druggable	Non-druggable
glf	O06934	Proposed research target	Non-druggable	Non-druggable
ldeR	P0A672	Proposed research target	Non-druggable	Non-druggable
ompA	P65593	Proposed research target	Non-druggable	Non-druggable
mshC	P67017	Proposed research target	Non-druggable	Non-druggable

accuracy for non-druggable proteins is better than that of druggable proteins. This probably results from the more diverse set of non-druggable proteins compared with that of druggable proteins, enabling SVMs to better recognize non-druggable proteins.

A total of 188 feature elements were used in deriving SVM models for predicting druggable proteins. In some classes, not all feature elements are essential, and their removal improves prediction performance [37]. Although higher numbers of descriptors are needed for representing the highly diverse druggable proteins, it is of interest to examine the extent to which feature reduction improves prediction performance. A rigorous feature selection method, recursive feature elimination (RFE) [38], was applied to the best-performing sampling set (fold-2) in Table 1. Implementation of RFE [38] is described in the supplementary material, from which 148 elements were selected and SVM performance was improved by 1%. These results confirm the need for a higher number of descriptors for predicting druggable proteins. The level of correlation in the 188-element descriptor set is relatively small, involving 21% of the elements. A reduction in these correlated elements has little effect on prediction performance, probably because they contribute an insubstantial level of noise. The full 188-element descriptor set is preferred because of its higher potential in covering novel druggable proteins.

TABLE 3

Statistics of predicted druggable proteins in different genomes by using an SVM prediction model

Genome	Number of predicted druggable proteins			Estimated druggable proteins or targets from other studies [Ref.]
	SVM model trained by both successfully commercialized and research targets	SVM model trained by successfully commercialized targets	SVM model trained by research targets	
H. sapiens	3379	662	2845	3051 [13]
S. cerevisiae	400	82	378	508 [13]
C. elegans	2687	398	2144	2267 [13]
D. melanogaster	1500	287	1238	1714 [13]
C. albicans	716	100	600	567 [42]
M. tuberculosis	845	105	732	333 [41], 354 [40]
H. influenzae	400	66	348	>40 [43], <478 [54]
H. pylori	277	43	237	594 [44]

Comparison with other methods

The performance of SVMs was compared with that of three other popular ML methods, probabilistic neural network (PNN), C4.5 decision tree and k nearest neighbor (kNN)], and with the sequence similarity method BLAST [39], based on the same fivefold cross-validation study. These methods are described in the supplementary material. The BLAST search was conducted between the drug-binding domains of each testing and training protein, to identify the training protein with the highest similarity score above a predetermined cut-off point. A testing protein is predicted as druggable or non-druggable when the identified training protein is druggable or non-druggable; and it is designated as non-druggable when no similar protein is identified. As shown in Table 1, the prediction accuracies for druggable and non-druggable proteins are in the range 52.5–69.3% and 75.9–83.3% for the ML methods, and 60.8–65.0% and 99.8–100% for the BLAST method, respectively. SVM outperforms all of these methods but shows a lower non-druggable prediction accuracy than BLAST, partly owing to the tentative designation of non-similar proteins as non-druggable. These proteins constitute 29.0–31.0% druggable and 81.6–85.5% non-druggable proteins in the testing sets, which favors non-druggable prediction performance by BLAST.

The performance of SVMs was further compared with that of the druggable domain affiliation (DDA) method. The DDA method has been tested on 35 published, successfully commercialized research and proposed research targets in the *Mycobacterium tuberculosis* genome [40]. DDA predicts druggable proteins by evaluating whether a member of the InterPro domain family of the studied protein is bound by a drug-like compound [40]. As shown in Table 2, 54% of the overall total of 35 targets and 64% of the 11 proposed research targets were predicted as druggable by SVMs, which is substantially better than the values of 31% and 45%, respectively, derived from DDA [40].

Evaluation of SVM-predicted druggable proteins in human, microbial and viral genomes

The numbers of SVM-predicted druggable proteins in the complete genomes of *Homo sapiens, Saccharomyces cerevisiae, Caenorhabditis elegans, Drosophila melanogaster, Candida albicans, Mycobacterium tuberculosis, Haemophilus influenzae and Helicobacter pylori* were compared with those predicted by other studies [13,41–44]. As shown in Table 3, the numbers of druggable proteins predicted by SVMs are highly consistent with the estimated numbers of targets or druggable proteins derived from other studies. Moreover, SVM-predicted druggable proteins in the genome of HIV-1, hepatitis C and influenza A H5N1 were compared with the known targets in these genomes. These viral genomes were selected because it is probable that all of their potential targets have been identified, owing to intensive research efforts directed at them [45]. The results are given in Table 4, Table 5 and Table 6, respectively. None of the encoded protein sequences in these genomes are in the SVM training and testing sets. There are four successfully commercialized and seven research targets in the HIV-1 genome, three research targets (one for the vaccine) in the hepatitis C genome and two successfully commercialized and two research targets in the influenza A H5N1 genome, respectively. SVMs correctly predicted all but one HIV research target, which was nucleocapsid.

To evaluate further whether SVMs predict druggable proteins rather than membership of certain Pfam families, Pfam family distribution of the predicted druggable proteins in the *H. sapiens* and *S. cerevisiae* genomes were analyzed. For the SVM model trained by using successfully commercialized and research targets, 16.4% and 31.8%, respectively, of the predicted druggable proteins in these genomes belong to Pfam families that contain no known successfully commercialized or research target. For the SVM model trained by using successfully commercialized targets, 34.1% and 53.7%, respectively, of the predicted druggable proteins in these genomes belong to Pfam families that contain no known successfully commercialized target, and 15.4% and 23.2%, respectively, of the predicted targets belong to Pfam families that contain no known successfully commercialized or research target. These results suggest that SVMs predict druggable proteins rather than membership to certain Pfam families.

Underlying difficulties in using SVMs for predicting druggable proteins

The performance of SVMs depends on the diversity of druggable and non-druggable proteins in a training dataset and the appropriate representation of the features of these proteins. The currently available datasets are not expected to be fully representative of all of the druggable and non-druggable proteins. Various

TABLE 4

Comparison of the known HIV-1 protein targets and the SVM-predicted druggable proteins in the NCBI[a] HIV-1 genome entry NC_001802

Protein	NCBI protein accession number	Target status	SVM prediction status
Gag-Pol	NP_057849.4	Non-target	Non-druggable
Gag-Pol transframe peptide	NP_787043.1	Non-target	Non-druggable
Pol	NP_789740.1	Non-target	Non-druggable
Protease	NP_705926.1	Successfully commercialized target	Druggable
Reverse transcriptase	NP_705927.1	Successfully commercialized target	Druggable
Reverse transcriptase p51 subunit	NP_789739.1	Research target	Druggable
Integrase	NP_705928.1	Research target	Druggable
Gag	NP_057850.1	Non-target	Non-druggable
Matrix	NP_579876.2	Non-target	Non-druggable
Capsid	NP_579880.1	Non-target	Non-druggable
p2	NP_579882.1	Non-target	Non-druggable
Nucleocapsid	NP_579881.1	Research target	Non-druggable
p1	NP_787042.1	Non-target	Non-druggable
p6	NP_579883.1	Non-target	Non-druggable
Vif	NP_057851.1	Research target	Druggable
Vpr	NP_057852.2	Non-target	Non-druggable
Tat	NP_057853.1	Successfully commercialized target	Druggable
Rev	NP_057854.1	Non-target	Non-druggable
Vpu	NP_057855.1	Non-target	Non-druggable
Envelope surface glycoprotein gp160	NP_057856.1	Research target	Druggable
Envelope signal peptide	NP_579893.2	Non-target	Non-druggable
Envelope surface glycoprotein gp120	NP_579894.2	Research target	Druggable
Envelope transmembrane glycoprotein gp41	NP_579895.1	Successfully commercialized target	Druggable
Nef	NP_057857.2	Research target	Druggable

[a] NCBI, National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/).

degrees of inadequate sampling are likely to affect the prediction accuracy of the developed SVM models. Discovery of novel targets and a deeper understanding of the characteristics of non-druggable proteins will enable further improvement of the prediction performance of SVMs.

In the available datasets, statistically, the number of druggable proteins is significantly smaller than that of non-druggable proteins. SVMs tend to push the hyperplane towards the side with a smaller number of samples [46], which leads to a reduced prediction accuracy for druggable proteins. However, it

TABLE 5

Comparison of the known hepatitis C protein targets and the SVM-predicted druggable proteins in the NCBI[a] hepatitis C virus genome entry NC_004102

Protein	NCBI protein accession number	Target status	SVM prediction status
Core protein	NP_751919.1	Non-target	Non-druggable
E1 protein	NP_751920.1	Non-target	Non-druggable
E2 protein	NP_751921.1	Vaccine research target	Druggable
p7 protein	NP_751922.1	Non-target	Non-druggable
NS2 protein	NP_751923.1	Non-target	Non-druggable
NS3 protease/helicase	NP_803144.1	Research target	Druggable
NS4A protein	NP_751925.1	Non-target	Non-druggable
NS4B protein	NP_751926.1	Non-target	Non-druggable
NS5A protein	NP_751927.1	Non-target	Non-druggable
NS5B RNA-dependent RNA polymerase	NP_751928.1	Research target	Druggable
Protein F	NP_671491.1	Non-target	Non-druggable

[a] NCBI, National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov).

TABLE 6

Comparison of the known H5N1 influenza A virus protein targets and the SVM-predicted druggable proteins in the eight NCBI[a] influenza A virus [A/Goose/Guangdong/1/96 (H5N1)] genome segment entries NC_007357, NC_007358, NC_007359, NC_007360, NC_007361, NC_007362, NC_007363 and NC_007364

Protein	NCBI protein accession number	Target status	SVM prediction status
Non-structural protein 2	YP_308672	Non-target	Non-druggable
Polymerase	YP_308664	Research target	Druggable
Polymerase	YP_308665	Non-target	Non-druggable
PB1-F2 protein	YP_473348	Non-target	Non-druggable
Polymerase	YP_308666	Non-target	Non-druggable
Nucleocapsid protein	YP_308667	Non-target	Non-druggable
Neuraminidase	YP_308668	Successful target	Druggable
Hemagglutinin	YP_308669	Research target	Druggable
HA1	YP_529486	Non-target	Non-druggable
HA2	YP_529487	Non-target	Non-druggable
Matrix protein 2	YP_308670	Successful target	Druggable
Matrix protein 1	YP_308671	Non-target	Non-druggable
Non-structural protein 1	YP_308673	Non-target	Non-druggable

[a] NCBI, National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov).

would be inappropriate simply to reduce the number of non-druggable proteins to match artificially that of druggable proteins, because this compromises the diversity needed for fully representing all types of non-druggable proteins. Instead, methods for readjusting the biased shift of hyperplanes are being explored [47].

A substantially higher number of descriptors are available than those used for predicting druggable proteins [29]. Selection of the most relevant subset of descriptors from the full set of descriptors is useful for improving the performance of SVMs [38]. Therefore, there is a need to explore different combinations of descriptors and to select an optimal set of descriptors by using feature selection methods [38]. Effort has been directed at improving the efficiency and speed of feature selection methods [48], which will enable a more extensive application of feature selection methods. Moreover, indiscriminate use of available descriptors, particularly overlapping and redundant ones, might introduce noise as well as extending the coverage of protein features. Investigation of cases of incorrectly predicted proteins has also suggested that the available descriptors might not always be sufficient for fully representing the properties of proteins [29]; this has prompted studies into developing new descriptors [25].

Perspectives

Statistical and proof-of-concept tests consistently show that SVMs are useful for facilitating the identification of druggable proteins. Rapid progress in genomics [49], structural genomics [50] and proteomics [51] is revolutionizing the process of target identification and drug development. In addition to the incorporation of newly discovered knowledge and information into SVMs and other *in-silico* methods, target identification can be further improved by the collective analysis of multiple sequence, structure, systems and physiological profiles [5,6,18,19,21], particularly sequence and functional similarity to known targets [13,14], drug-binding domain family affiliation [7,13], geometric and energetic features of protein structures [15,16], ligand–protein inverse docking [52] and systems-related properties [18,21]. These methods might potentially be developed into useful tools for facilitating the identification of novel targets. These developments, combined with advances in the molecular understanding of disease processes [53], have opened opportunities for discovering new and novel targets.

Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.drudis.2007.02.015.

References

1 Ohlstein, E.H. *et al.* (2000) Drug discovery in the next millennium. *Annu. Rev. Pharmacol. Toxicol.* 40, 177–191

2 Drews, J. (1997) Strategic choices facing the pharmaceutical industry: a case for innovation. *Drug Discov. Today* 2, 72–78

3 Walke, D.W. *et al.* (2001) *In vivo* drug target discovery: identifying the best targets from the genome. *Curr. Opin. Biotechnol.* 12, 626–631

4 Ilag, L.L. *et al.* (2002) Emerging high-throughput drug target validation technologies. *Drug Discov. Today* 7 (18 Suppl.), S136–S142

5 Lindsay, M.A. (2005) Finding new drug targets in the 21st century. *Drug Discov. Today* 10, 1683–1687

6 Sams-Dodd, F. (2005) Target-based drug discovery: is something wrong? *Drug Discov. Today* 10, 139–147

7 Kramer, R. and Cohen, D. (2004) Functional genomics to new drug targets. *Nat. Rev. Drug Discov.* 3, 965–972

8 Ryan, T.E. and Patterson, S.D. (2002) Proteomics: drug target discovery on an industrial scale. *Trends Biotechnol.* 20 (12 Suppl.), S45–S51

9 Lindsay, M.A. (2003) Target discovery. *Nat. Rev. Drug Discov.* 2, 831–838
10 Nicolette, C.A. and Miller, G.A. (2003) The identification of clinically relevant markers and therapeutic targets. *Drug Discov. Today* 8, 31–38
11 Jackson, P.D. and Harrington, J.J. (2005) High-throughput target discovery using cell-based genetics. *Drug Discov. Today* 10, 53–60
12 Austen, M. and Dohrmann, C. (2005) Phenotype-first screening for the identification of novel drug targets. *Drug Discov. Today* 10, 275–282
13 Hopkins, A.L. and Groom, C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov.* 1, 727–730
14 Wang, S. *et al.* (2004) Tools for target identification and validation. *Curr. Opin. Chem. Biol.* 8, 371–377
15 Hajduk, P.J. *et al.* (2005) Druggability indices for protein targets derived from NMR-based screening data. *J. Med. Chem.* 48, 2518–2525
16 Hajduk, P.J. *et al.* (2005) Predicting protein druggability. *Drug Discov. Today* 10, 1675–1682
17 Booth, B. and Zemmel, R. (2004) Prospects for productivity. *Nat. Rev. Drug Discov.* 3, 451–456
18 Zheng, C. *et al.* (2006) Progress and problems in the exploration of therapeutic targets. *Drug Discov. Today* 11, 412–420
19 Hardy, L.W. and Peet, N.P. (2004) The multiple orthogonal tools approach to define molecular causation in the validation of druggable targets. *Drug Discov. Today* 9, 117–126
20 Han, L.Y. *et al.* (2004) Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucleic Acids Res.* 32, 6437–6444
21 Zheng, C.J. *et al.* (2006) Therapeutic targets: progress of their exploration and investigation of their characteristics. *Pharmacol. Rev.* 58, 259–279
22 Bao, L. and Sun, Z. (2002) Identifying genes related to drug anticancer mechanisms using support vector machine. *FEBS Lett.* 521, 109–114
23 Karchin, R. *et al.* (2002) Classifying G-protein coupled receptors with support vector machines. *Bioinformatics* 18, 147–159
24 Bhasin, M. and Raghava, G.P. (2004) GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Res.* 32 (Web Server issue), W383–W389
25 Bhardwaj, N. *et al.* (2005) Kernel-based machine learning protocol for predicting DNA-binding proteins. *Nucleic Acids Res.* 33, 6486–6493
26 Lin, H.H. *et al.* (2006) Prediction of transporter family from protein sequence by support vector machine approach. *Proteins* 62, 218–231
27 Bhasin, M. and Raghava, G.P. (2004) Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J. Biol. Chem.* 279, 23262–23266
28 Cai, C.Z. *et al.* (2004) Enzyme family classification by support vector machines. *Proteins* 55, 66–76
29 Han, L. *et al.* (2006) Recent progresses in the application of machine learning approach for predicting protein functional class independent of sequence similarity. *Proteomics* 6, 4023–4037
30 Baldi, P. *et al.* (2000) Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 16, 412–424
31 Chen, X. *et al.* (2002) TTD: Therapeutic Target Database. *Nucleic Acids Res.* 30, 412–415
32 Chantry, D. (2003) G protein-coupled receptors: from ligand identification to drug targets. 14–16 October 2002, San Diego, CA, USA. *Expert Opin. Emerg. Drugs* 8, 273–276
33 Finn, R.D. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.* 34 (Database issue), D247–D251
34 Li, Z.R. *et al.* (2006) PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.* 34 (Web Server issue), W32–W37
35 Gasteiger, E. *et al.* (2005) Protein identification and analysis tools on the ExPASy server. In *The Proteomics Protocols Handbook* (John, M.W., ed.), pp. 571–607, Humana Press
36 Cai, C.Z. *et al.* (2003) SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.* 31, 3692–3697
37 Ding, C.H. and Dubchak, I. (2001) Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* 17, 349–358
38 Xue, Y. *et al.* (2004) 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
39 McGinnis, S. and Madden, T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.* 32 (Web Server issue), W20–W25
40 Hasan, S. *et al.* (2006) Prioritizing genomic drug targets in pathogens: application to *Mycobacterium tuberculosis*. *PLoS Comput Biol* 2, e61
41 Cole, S.T. (2002) Comparative mycobacterial genomics as a tool for drug target and antigen discovery. *Eur. Respir. J. Suppl.* 36, 78s–86s
42 Roemer, T. *et al.* (2003) Large-scale essential gene identification in *Candida albicans* and applications to antifungal drug discovery. *Mol. Microbiol.* 50, 167–181
43 Huynen, M.A. *et al.* (1997) Differential genome display. *Trends Genet.* 13, 389–390
44 Huynen, M. *et al.* (1998) Differential genome analysis applied to the species-specific features of *Helicobacter pylori*. *FEBS Lett.* 426, 1–5
45 Turpin, J.A. (2003) The next generation of HIV/AIDS drugs: novel and developmental antiHIV drugs and targets. *Expert Rev. Anti Infect. Ther.* 1, 97–128
46 Veropoulos, K. *et al.* (1999) Controlling the sensitivity of support vector machines. In *Proceedings of the International Joint Conference on Artificial Intelligence (UCAI99)* (Dean, T., ed.), In pp. 55–60, Morgan Kaufmann
47 Brown, M.P. *et al.* (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. U. S. A.* 97, 262–267
48 Furlanello, C. *et al.* (2003) An accelerated procedure for recursive feature ranking on microarray data. *Neural Netw.* 16, 641–648
49 Debouck, C. and Metcalf, B. (2000) The impact of genomics on drug discovery. *Annu. Rev. Pharmacol. Toxicol.* 40, 193–207
50 Sali, A. (1998) 100,000 protein structures for the biologist. *Nat. Struct. Biol.* 5, 1029–1032
51 Dove, A. (1999) Proteomics: translating genomics into products? *Nat. Biotechnol.* 17, 233–236
52 Chen, Y.Z. and Zhi, D.G. (2001) Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule. *Proteins* 43, 217–226
53 Macdonald, I.A. (2000) Obesity: are we any closer to identifying causes and effective treatments? *Trends Pharmacol. Sci.* 21, 334–336
54 Akerley, B.J. *et al.* (2002) A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proc. Natl. Acad. Sci. U. S. A.* 99, 966–971