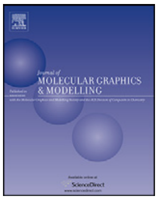
Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

journal homepage: www.elsevier.com/locate/JMGM



Data mining PubChem using a support vector machine with the Signature molecular descriptor: Classification of factor XIa inhibitors

Derick C. Weis [a], Donald P. Visco Jr. [a,*], Jean-Loup Faulon [b]

[a] Department of Chemical Engineering, Tennessee Technological University, Box 5013, Cookeville, TN 38505, United States
[b] Sandia National Laboratories, Computational Biosciences Department, P.O. Box 5800, Albuquerque, NM 87185, United States

ARTICLE INFO

Article history:
Received 9 July 2008
Received in revised form 19 August 2008
Accepted 20 August 2008
Available online 27 August 2008

Keywords:
De novo
Signature
Factor XIa
HTS

ABSTRACT

The amount of high-throughput screening (HTS) data readily available has significantly increased because of the PubChem project (http://pubchem.ncbi.nlm.nih.gov/). There is considerable opportunity for data mining of small molecules for a variety of biological systems using cheminformatic tools and the resources available through PubChem. In this work, we trained a support vector machine (SVM) classifier using the Signature molecular descriptor on factor XIa inhibitor HTS data. The optimal number of Signatures was selected by implementing a feature selection algorithm of highly correlated clusters. Our method included an improvement that allowed clusters to work together for accuracy improvement, where previous methods have scored clusters on an individual basis. The resulting model had a 10-fold cross-validation accuracy of 89%, and additional validation was provided by two independent test sets. We applied the SVM to rapidly predict activity for approximately 12 million compounds also deposited in PubChem. Confidence in these predictions was assessed by considering the number of Signatures within the training set range for a given compound, defined as the overlap metric. To further evaluate compounds identified as active by the SVM, docking studies were performed using AutoDock. A focused database of compounds predicted to be active was obtained with several of the compounds appreciably dissimilar to those used in training the SVM. This focused database is suitable for further study. The data mining technique presented here is not specific to factor XIa inhibitors, and could be applied to other bioassays in PubChem where one is looking to expand the search for small molecules as chemical probes.

1. Introduction

High-throughput screening (HTS) is commonly used in drug discovery to find new lead compounds by physically screening large libraries of compounds against a given biological target. In the past, this technique was primarily available only to the pharmaceutical industry and not to academic researchers. The Molecular Libraries Initiative (MLI) [1], part of the NIH Roadmap for Medical Research [2], sought to increase the use of small molecules as chemical probes for basic research. The MLI established the Molecular Libraries Screening Center Network (MLSCN) [3] which currently consists of 10 centers located across the country to perform HTS for assays submitted by the research community. Results from all screens are deposited in a public archive known as PubChem that is comprised of three databases: PCSubstance, PCCompound and PCBioAssay [4]. PCSubstance currently contains more than 38 million compounds, 17 million of which are unique and are contained in PCCompound. Results from more than 800 different HTS experiments can be found in PCBioAssay, and are linked to the corresponding compounds in PCSubstance and PCCompound. The public is welcome to download or deposit compounds or bioassay data into PubChem for free. In addition, searches can be performed to find exact or similar compounds by text entry, SMILES, molecular formula, drawing structures with an online tool, or various chemical structure file formats [4].

Owing to PubChem, there is now a tremendous amount of data readily available on how small molecules affect biological systems. In order to efficiently analyze and interpret this vast and growing database, cheminformatic tools are necessary. To meet these goals, Oprea et al. discussed the need to combine cheminformatic tools with currently available bioinformatic and cheminformatic databases. Such an approach would provide a means for a methodical understanding of how small molecules affect biological systems [5]. According to these authors, research performed in this area will create a new field called "systems chemical biology".

Xie and Chen data mined PubChem using a cell-based partition algorithm to create a representative subset of compounds from

* Corresponding author. Tel.: +1 931 372 3606.
E-mail addresses: dcweis21@tntech.edu (D.C. Weis), dvisco@tntech.edu (D.P. Visco Jr.), jfaulon@sandia.gov (J.-L. Faulon).

1093-3263/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2008.08.004

PubChem [6]. The total number of compounds was reduced to 540k from approximately 5.3 million to make *in silico* or *in vitro* screening of PubChem more manageable. The level of annotation information available from PubChem was studied by using the 2.5 million compound database at the Genomics Institute of the Novartis Research Foundation (GNF) as a basis for comparison [7]. These authors found that 32% of the GNF compounds could be linked to other databases by PubChem. Rosania et al. reviewed the current cheminformatic tools available to mine biomedical knowledge, and the challenges involved to implement these tools for users who do not have extensive training in cheminformatics [8]. Ingsriswang and Pacharawongsakda developed a web based tool known as sMOL Explorer [9]. The main functions of the program include data management, performing statistical analysis, and data mining. Fontaine et al. used alignment-recycling to reduce the CPU time required to perform 3D shape similarity searches for large databases [10]. Finally, Li et al. utilized bioassay data from PubChem as an independent test set for a support vector machine (SVM) trained to classify human Ether-a-go-go Related Gene potassium channel inhibitors, and obtained about 73% accuracy [11].

SVM is a supervised machine-learning method used to classify data as input vectors by creating an optimal separating hyperplane [12]. Training a SVM has become a popular technique in cheminformatic applications. It has been used to predict various properties for small molecules including biological activity [13–19], metabolism by cytochrome P450 [20–22], toxicity [23,24], and blood–brain barrier penetration [25]. As SVMs have outperformed other statistical learning methods [14,25], we will use it in this work.

For our study, the input vectors used in the SVM are graph-based representations of molecules known as Signature [26], where chemical information is encoded by canonical subgraphs [27]. Signature has previously been applied with a SVM to predict both protein–protein [28] and drug–target [19] interactions. The number of descriptors (referred to as features when used in statistical learning methods) compared to the number of observations is an important consideration to avoid overfitting. Filters and wrapper methods are the main procedures used for feature selection. Filtering feature selection methods apply a metric to initially rank features independent of SVM prior to training. Wrapper algorithms conduct training/testing phases with SVM where the goal is to choose important features that improve performance.

An example of a system studied via HTS in the PCBioAssay database is screening for inhibitors of factor XIa. Factor XI is a serine protease expressed as a zymogen [29], and then converted to an activated form (factor XIa) [30] involved in the amplification phase of the coagulation cascade. It is an interesting therapeutic target because it has the potential for development of novel antithrombotic drugs to replace conventional ones like heparin and warfarin [31]. Because factor XIa is involved in the amplification phase of coagulation and not clotting, specific inhibitors could treat thrombosis without the risk of severe bleeding present for inhibiting other coagulation proteases. A primary screening was performed at the Penn Center for Molecular Discovery (PCMD) on approximately 200,000 compounds at a single concentration of 5 μM and the data reported in the PCBioAssay database as assay identification (AID) 798 [32]. Any compound providing at least 40% inhibition was determined to be active, and screened again in a dose response confirmatory assay for IC_{50} determination in AID 846 [33]. Jin et al. reported the first crystal structure of recombinant factor XI for structure-based design of selective inhibitors [34]. Several recent studies have focused on structure-based design and synthesis of factor XIa inhibitors [35–37]. However, owing to their ability to process large amounts of information, cheminformatic tools can also be used to advance the search for inhibitors of factor XIa.

In this work, we present a method to data mine the *entire* PCCompound database using SVM to suggest compounds for additional HTS of factor XIa inhibitors with a higher probability of being active than random selection. Details on Signature and selecting an appropriate number of Signatures (features) for SVM are provided in Section 2. Section 3 compares a filtering approach feature selection to a wrapper algorithm that recursively eliminates clusters of highly correlated features. A subset of Signatures was chosen to train a SVM on the confirmatory HTS data (AID 846). Inactive compounds from the primary screen (AID 798) were applied as a test set to evaluate the model for false positive predictions. Next, approximately 12 million compounds from the PCCompound database were screened to form a focused database of predicted factor XIa inhibitors. All computations were performed on a 2.8 GHz Pentium 4 Xeon processor.

2. Methods

2.1. Signature

The Signature molecular descriptor has previously been described in detail [26,27,38–41]. A brief introduction and review is provided here for the assistance of the reader. Signature is a topological descriptor providing an efficient method to encode the local neighborhood of a molecule. The size of the local neighborhood is controlled by a parameter called the Signature height, *h*. An atomic signature consists of a subgraph starting at a root atom, but includes all atoms and bonds out to the predefined height *h*. The atomic signature of height *h* is formally given as ${}^{h}\sigma_{G(x)}$ for the root atom *x* of the 2D graph $G = (\upsilon, E)$, where υ, E is vertex (atom) set and edge (bond) set respectively. The molecular signature is formally given as

$$ {}^{h}\sigma(G) = \sum_{x \in V}^{h} \sigma_{G}(x) \tag{1} $$

where ${}^{h}\sigma(G)$ is a vector of occurrences for all the unique atomic signatures.

There are many options available when choosing a cheminformatic descriptor. Signature was selected for this study because it has previously exhibited a low degeneracy [39] and has produced predictive models [40]. The degeneracy of Signature can actually be fine-tuned by adjusting the height selected. A low degeneracy is an important quality for a molecular descriptor to distinguish between active and inactive compounds. For example, consider molecular weight as a descriptor for classification. It is possible for two compounds to have the same molecular weight, but very different activity due to chemical structure. Thus, molecular weight would not be an appropriate descriptor for classification of activity. Compared to other popular molecular descriptors, Signature was shown to be the least degenerate on an assortment of molecular series including alkanes, alcohols, fullerenes, and peptides [39]. The ability to create a useful model is another consideration for selecting a descriptor. A QSAR created with Signature using a forward-stepping multiple linear regression (MLR) compared favorably with descriptors from the commercially available program Molcon-Z [42] on a small set (~100 compounds) of HIV-1 protease inhibitors, and on a large set (~10,000 compounds) with octanol/water partitioning coefficient [40]. Martin et al. first combined Signature with SVM and defined the Signature kernel for variable length amino acid sequences

$$ k(\mathrm{A}, \mathrm{B}) = s(\mathrm{A}) \cdot s(\mathrm{B}) \tag{2} $$

where $s(\mathrm{A})$ and $s(\mathrm{B})$ is the Signature of amino acid sequence A and B respectively [28]. Protein sequences were encoded with Signature to train a SVM for prediction of protein–protein interactions. A unique kernel was developed using Signature products to describe pairs of interacting proteins instead of a single protein sequence. The technique was extended to also describe enzyme–metabolite and drug–target pairs [19]. Enzymes catalyzing reactions and drugs binding to a specific target were predicted without including that information for SVM learning.

2.2. Support vector machines

SVMs [12] are classifiers that use an optimal separating hyperplane and have been applied in a variety of cheminformatic applications. Specifically for biological activity, a study was performed using Molecular Operating Environment and topological pharmacophore (CATS) descriptors to characterize focused libraries of kinase, factor Xa, and thrombin inhibitors [16]. Plewczynski et al. used atom pair descriptors to classify compounds from the MDL drug data report to work with cyclooxygenase-2, dihydrofolate reductase, thrombin, HIV-reverse transcriptase and antagonists of the estrogen receptor [13]. Jorissen and Gilson focused on enrichment of actives instead of classification by modifying SVM to rank molecules [15]. Glick et al. used several statistical methods including Laplacian-modified naïve Bayes, recursive partitioning, and SVM for enrichment using HTS data [14]. Noise was deliberately introduced as false positives and false negatives to evaluate the ability of the model to perform under such conditions. SVM proved to be superior at finding the compounds and scaffolds in the top 1%.

To explain SVMs, assume data are presented as pairs $\{(x_i, y_i)\} \subset R^n\{\pm 1\}$, which means that data belongs to only two classes with labels +1 or −1. In our case, +1 refers to active compounds, and −1 represents inactive compounds. The specific definition for active and inactive can vary depending on the system and researcher. Using the pair notation, SVMs are given in the form

$$f(x) = \sum\nolimits_i \alpha_i y_i k(x_i, x) + b \tag{3}$$

where f: $R^n \rightarrow R$ is a decision function for classification. If $f(x)$ is greater than some threshold t, then x belongs to the class +1, if not x belongs to the class −1. The constants α_i and b are acquired by solving the quadratic programming problem. The constant α_i is zero for all observations except the important borderline cases, which are known as the support vectors. A kernel function k: $R^n \times R \rightarrow R$ is a dot product in some vector space that can efficiently transform input data to higher dimensions if desired. We use the Signature kernel described in Eq. (2) for chemicals instead of amino acid sequences. The vector space will be the molecular signature made up of all unique atomic signatures from heights 0, 1 and 2. For additional details on SVM, the tutorial by Burgess is recommended [43]. The SVMs in this work were generated by the SVM$^{\text{light}}$ algorithm [44]. Ideally the data would be classified with SVM without having any errors, but this is not realistically possible. The user defined cost parameter C controls the tradeoff between allowing for some misclassification and the margin of the optimal separating hyperplane. For this problem, we evaluated cost parameters that ranged several orders of magnitude using a search strategy following an approach used elsewhere. [15] Ultimately, we found that a value of 1.0 balanced the tradeoff and, accordingly, was applied here in the SVM$^{\text{light}}$ program.

2.3. Feature selection

Even though SVM is less susceptible to overfitting compared to other machine learning methods, [43] selecting a subset of relevant features (Signatures) is necessary to increase predictive capability. There are two main categories of feature selection strategies known as filter and wrapper methods. A filtering method ranks individual features by a defined metric completely independent from SVM. On the other hand, a wrapper method selects features to incorporate into the model by working with the SVM during training/testing steps where the goal is to optimize some objective function.

The advantage of filter methods is computational efficiency, with the downside being that each feature is treated separately. There is an implicit orthogonal assumption here that rarely holds because features are usually correlated. Filtering methods attempt to identify those features that discriminate the most between the two classes available for classification. The coefficient ω_i is an example of a filtering metric defined by Golub as

$$\omega_i = \frac{\mu_i(+) - \mu_i(-)}{\sigma_i(+) + (\sigma_i(-)} \tag{4}$$

where μ_{i} and σ_{i} are the mean and standard deviation for features in the (+) or (−) class, respectively [45]. Large positive values of ω_i signify a relationship to the (+) class, while large negative values correspond to the (−) class. The ω_i coefficient was used in a bioinformatics application to rank gene expression values for gene i for leukemia classification. An equal number of genes from each class were selected in this method. The absolute value of ω_i has also been suggested as a ranking criteria to classify ovarian cancer tissue using gene expression data [46]. In our study, we have chosen Furey's use of the ω_i to implement the filter method for feature selection.

In contrast to the filter method, the wrapper method uses an iterative approach to enhance SVM performance. One such comprehensive work on wrapper methods was performed by Kohavi and John [47], who showed accuracy improvement using hill-climbing and best-first search to select feature subsets for decision trees and Naïve-Bayes algorithms. Recently, Yousef et al. [48] combined K-means clustering with SVM for feature selection in an algorithm known as recursive cluster elimination for work on gene expression. Using this approach, the clusters were scored individually which resulted in a fast algorithm, but did not allow the clusters to work together for improved accuracy. In our work, we have utilized a similar approach as Yousef, but instead apply a more rigorous algorithm which removes a cluster to train the SVM on all remaining clusters in each step in an attempt to arrive at an improved model. We describe this approach in more detail below.

Prior to SVM training, features are grouped into K mutually exclusive clusters based on the Pearson correlation coefficient using Clusteran [49]. Features are first randomly assigned to one of the K clusters, then iteratively compared to other clusters and moved, if necessary, to best group the correlated features. The iteration continues until a stable division of the specified number of K clusters is achieved. Initially all K clusters are included for SVM training. Each cluster is then sequentially dropped, and the remaining clusters are used for SVM training where accuracy is assessed by 10-fold cross-validation as described below. The cluster that provided the highest accuracy when dropped was permanently removed from consideration. The process is then repeated on the surviving $K - 1$ clusters to ultimately find an optimal subset of clusters that maximize SVM accuracy. When more than one cluster provided the same accuracy, the absolute value from the decision function in Eq. (3) was summed for all misclassified compounds, and applied as a tiebreaker.

2.4. Cross-validation

Evaluating the performance of the SVM was carried out by using X-fold cross-validation. The training set is first divided into X

subsets containing an equal number of compounds, where each compound appears only once. One subset of compounds is withheld for testing while the remaining subsets are used for SVM training. The process is repeated for all X subsets to provide predictions of all compounds in the dataset when they are not included for training purposes. In this work, we used 10-fold cross-validation for feature selection.

The predictions from cross-validation were assigned the notation TP, FP, TN or FN for true positives, false positives, true negatives and false negatives, respectively. The statistics accuracy, sensitivity, specificity and precision were averaged over the X-fold subsets and computed as defined below.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \tag{5}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{7}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

Normally the threshold, t, is set to zero for SVM causing any prediction greater than zero to be classified as active, and any prediction less than zero to be inactive. Changing the threshold t alters the values of TP, FP, TN and FN for predictions with a decision value close to zero, which influences SVM performance. Integrating the area under the receiver–operator characteristic (ROC) curve for each X-fold subset provides an additional averaged statistic for performance evaluation. Varying the threshold t over the predicted SVM decision function values from a given X-fold subset and plotting the TP rate (sensitivity) verses the FP rate (1-specificity) creates the ROC curve. We investigate raising the threshold t value to screen the PCCompound database to increase precision. Some TP compounds are sacrificed because of this, but when an active prediction is made, there is more confidence in this selection.

2.5. Construction of training and test set

Working with HTS data using SVM can be a challenging task for two major reasons. First, the number of inactive compounds almost always exceeds the number of active compounds in primary screens by a significant amount, causing highly unbalanced datasets. Machine learning methods typically do not perform well on unbalanced sets because of a tendency to classify all observations the same as the majority class [50]. For example, if only 2% of the compounds were active, classifying all compounds as inactive would yield an accuracy of 98%, but not be a useful model. The second issue has to do with the quality of data obtained from HTS being noisy primarily from false positives [51]. For these reasons we decided to train on the confirmatory assay (AID 846) because it is relatively balanced and false positives are not as likely because the IC_{50} determination was performed in triplicate. The assay depositor reported a compound as active if $IC_{50} < 50\ \mu M$ was obtained in all three IC_{50} determinations, inconclusive if $IC_{50} < 50\ \mu M$ in only one or two determinations, and inactive for $IC_{50} > 50\ \mu M$ [33]. For this work, we reduced the active classification to 5 μM resulting in 47 active and 68 inactive compounds. Changing the activity classification allowed inactive compounds from the primary screen (AID 798) to be directly applied as a large test set for false positive predictions because these compounds had less than 40% inhibition from a single measurement at 5 μM.

2.6. Overlap metric

Since Signatures are obtained unique to a given data set, not all of the Signatures used to train the SVM (using AID 846) necessarily are contained in either the AID 798 test set and/or the PCCompound database. It stands to reason then that (in general) predictions obtained for a compound using the SVM would be more reliable when there is more overlap between the Signatures found in the compound and that set used for the SVM. To quantitatively evaluate confidence in the prediction obtained for a given compound, we define an overlap metric, Ω, as

$$\Omega = \frac{x_{\text{min-max}}}{x_{\text{m}}} \tag{9}$$

where x_{m} is the total number of Signatures in a compound, and $x_{\text{min-max}}$ is the total number of Signatures common with the training set within the minimum/maximum occurrence range. We include this latter stipulation on occurrence range to surface extrapolation effects for individual Signatures. The values of Ω will range from 0 to 1 with predictions on the compounds with higher Ω values considered to be more reliable.

3. Results

The first step in the study was to determine the SVM using Signature for AID 846. All totaled, AID 846 (which contained 115 compounds) produced 865 unique height 0, 1 and 2 atomic signatures. Since the problem size scales with the number of Signatures, we reduced the Signature database by more than half (to 411 Signatures) when only including Signatures that have occurred in at least two compounds. Once this was accomplished, the next step was to perform the feature selection using our two techniques: the filter method and the wrapper method.

Recall that for the filter method, we use all 411 height 0, 1 and 2 atomic signatures in a reverse removal manner where the lowest ranked Signatures using the ω_i metric are dropped and performance is evaluated. Using this naïve filtering approach, a maximum accuracy of 80% is obtained (as seen in Fig. 1a) in under 10 min of CPU time. Note that without any feature selection, using all of the 411 Signatures results in an SVM with an accuracy of only 56%.

Once this was complete, we employed the wrapper technique which uses K-means clustering on the set of 411 Signatures. This approach increased performance to 89% accuracy, but at an expense of approximately four days of CPU time. The significant difference in CPU time between the two methods was because of the nested loops in the clustering method. The filtering approach just ranks the Signatures using the ω_i metric while the wrapper technique employs nested subsets. In the latter approach, the subsets were obtained starting with all Signatures and then dropping four at a time to perform the 10-fold cross-validation. Note that in our opinion even though the wrapper technique was nearly 600 times as computationally expensive as the filter technique, once the model has been obtained CPU time is no longer as important an issue. Thus the almost 10% increase in accuracy warrants the extra computational and time burden.

As can be seen from Fig. 1b, the highest cross-validation accuracy for the wrapper method occurs with 22 clusters (involving a total of 105 Signatures) and, thus, we choose this subset for the final SVM training. Notice that with only a small number of clusters (re: Signatures), there is not sufficient information available to construct a predictive model. On the other hand, when too many clusters (re: Signatures) are included, there is clearly overfitting. Additional statistics from the optimal Signature subset is provided in Table 1 for both 4 and 10-fold cross-

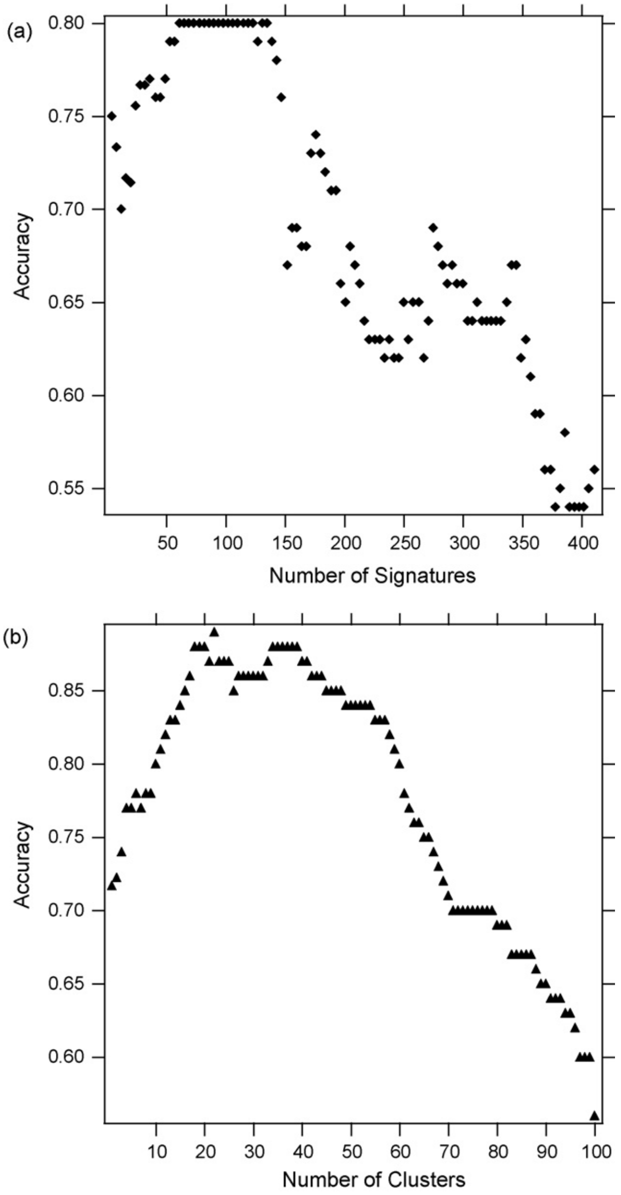


Fig. 1. Feature selection improves SVM performance evaluated by 10-fold cross-validation. A filtering approach (a) was compared to a clustering (b) technique.

validation. All subsequent results to both evaluate this SVM relative to AID 798 and predict potential actives in PCCompound are based on this SVM.

3.1. False positive testing

The 218,416 inactive compounds from AID 798 provided a large test set to evaluate for false positive predictions. The test set was initially filtered by an increasing Ω value, and then predictions were performed for a threshold t of 0, 1, and 2. So, for example, when $\Omega > 0$ (which contains all compounds in this set as they are all identified as inactive), increasing the value of the threshold results in an improved classification (see Table 2). The specificity metric, which is the ratio of true negatives to the sum of true negatives and false positives, is provided in the square brackets. For a given threshold value, as the overlap metric (Ω) increases, the number of compounds in the AID 798 satisfying this criterion decreases while the specificity increases. Ultimately, when the threshold is high ($t > 2$) and the overlap is maximum ($\Omega = 1$), the SVM perfectly classifies the 1442 compounds as inactive. Thus, the next step is to use this SVM to predict actives for factor XIa in the entire PCCompound database.

Table 1
Prediction statistics for training SVM on factor XIa inhibitor data (47 active and 68 inactive) with 22 clusters (105 Signatures)

X-fold	Accuracy	AUC (ROC)	Precision	Sensitivity	Specificity
10	0.8900	0.8856	0.8500	0.9000	0.8833
4	0.8125	0.8676	0.7568	0.7727	0.8382

3.2. Data mining PubChem

Now that we have created our SVM for factor XIa, the next step is to classify all of the compounds available in the PCCompound database, minus those from AID 798. To aid the reader in understanding the process as a whole, we provide Fig. 2 which is a flowsheet of the general algorithm employed in this research. Ultimately, our goal is to arrive at a ‘‘focused database’’ of compounds which are predicted as actives against factor XIa from the model. Such information could inform future HTS work on factor XIa.

At the time of downloading, the entire PCCompound database consisted of 11,946,913 unique chemical structures. Note that both training and test sets used in this work are contained within this database. All aromatic bonds were encoded using the Marvin Beans [52] software package since structures in the PCCompound database are reported using single and double bond notation. Code from previously developed algorithms [27] then translated the compounds to atomic signature heights of 0, 1 and 2. The combined CPU time to run the Marvin Beans package and to obtain all of the Signatures was approximately 1.5 days divided over for 4 machines for a total of about 6 days.

In Table 3 we provide the results in a format similar to that for the test set analysis of AID 798 with, of course, no metric for accessing the predictive ability of the SVM against these nearly 12 million compounds. However, we do report a predicted percent active per compounds evaluated that satisfy the specified overlap metric in square brackets. While no information is available for the activity of these compounds against factor XIa in PubChem, this study does identify potentially useful focused databases which can, ultimately, be screened to determine actives. We report all of the information available in Table 3 in supplemental documentation archived electronically with this paper. Examining the magnitude of the decision function for a variety of compounds above, it was observed that the spread of values were much larger when the overlap metric was smaller. Accordingly, we plotted a histogram of this information for three different overlap metrics in Fig. 3. While training the SVM on the 115 compounds in AID 846, the maximum value for the magnitude of the decision function was 2.10. It would stand to reason, therefore, that maximal predicted magnitudes should be reasonably close to this value. As can be seen in Fig. 3, this becomes increasingly true as $\Omega \rightarrow 1$. Thus, we would have more confidence in those values from Table 3 which are down and to the right. It is our recommendation, therefore, that future HTS screens on factor XIa start with compounds in that area of the table.

As a way to explore some of the compounds in this region of Table 3, we look at all 1300 compounds that have a magnitude of the decision function (re: t) greater than zero. Next, we calculated the set-theoretic Tanimoto coefficient (TC) between these 1300 compounds and all 115 compounds from AID 846 using all height 0–2 Signatures. TC is commonly used as a metric to assess the similarity between chemicals, where the values range from 0 to 1 [53]. Those chemicals with more similarity have a TC closer to 1.

Table 2
Test set prediction statistics from inactive factor XIa inhibitors in AID 798

t	$\Omega > 0$ (n = 218,416)	$\Omega > 0.6$ (n = 213,104)	$\Omega > 0.7$ (n = 196,352)	$\Omega > 0.8$ (n = 135,021)	$\Omega > 0.9$ (n = 37,655)	$\Omega = 1.0$ (n = 1442)
>0	193,762 [0.8871]	189,648 [0.8899]	176,062 [0.8967]	123,415 [0.9140]	35,226 [0.9355]	1359 [0.9424]
>1	211,282 [0.9673]	206,533 [0.9692]	190,957 [0.9725]	132,352 [0.9802]	37,103 [0.9853]	1424 [0.9875]
>2	216,210 [0.9899]	211,157 [0.9909]	194,896 [0.9926]	134,414 [0.9955]	37,590 [0.9983]	1442 [1.0000]

The numbers in parenthesis indicate the number of compounds from AID 798 which were above the various Ω values while the table entries indicate how many of those compounds were identified as inactive. The square brackets indicate specificity.

The incentive for determining this similarity measure is to evaluate the level of diversity amongst these 1300 compounds to those from the training set. Potentially, compounds which score low on this similarity measure (meaning that are most dissimilar to the training set) might constitute those non-intuitive selections of active inhibitors against Factor XIa. Accordingly, we report in Table 4 a sample of 12 compounds from the 1300 which have $t > 1$. For each compound, we also provide the value for the maximum Tanimoto coefficient that this compound has with the 115 compounds from AID 846. While some compounds, like CID2658123 (not shown), are just a small perturbation from one of the training set compounds (in this case, an additional –O-CH_3 group on the benzene ring), others (like CID 6104501) bare only marginal resemblance to anything in the training set.

3.3. Docking of PubChem compounds

Though the procedure up to this point provides a cheminformatics identification of potential factor XIa inhibitors through an SVM model, we wanted to provide an additional test on the validity of our proposed technique. To this end, we chose to perform a docking study on a subset of compounds that our technique has identified as being inhibitors of factor XIa in order to determine binding energies. In particular, we used AutoDock version 4.0.1 [54]. The AutoDock tools GUI was used to prepared the factor XIa crystal structure (PDB 1zpc) [35] in complex with a ligand. Crystallographic waters were removed, polar hydrogens were added, and a 50 × 50 × 50 grid box with 0.375 Å spacing was specified and centered on the active site. The Lamarckian genetic algorithm option for ligand conformational searching was applied for all compounds in this work with the following docking parameters: 100 runs, population size of 150, random starting point, 27,000 generations, and 25,000,000 energy evaluations.

To provide a basis for evaluating the binding energy of compounds identified by SVM, both known active and inactive compounds were initially docked. The 47 compounds from AID 846 classified as active ($IC_{50} < 5.0\ \mu M$) for SVM training all had a minimum binding energy ranging from −5.59 to −9.15 kcal/mol. In contrast, 25 compounds with less than 5% inhibition from AID 798 were randomly selected and found to have a minimum binding energy varying from −0.12 to 0.96 kcal/mol. The bi-modal distribution seen here provides motivation for additional consideration of compounds possessing strong binding energies docked to factor XIa. While a large, negative binding energy does not guarantee success as an inhibitor, it does serve as an additional tool for evaluation purposes. For example, Li et al. virtually screened the National Cancer Institute diversity set for AICAR transformylase inhibitors using AutoDock [55]. Approximately 50% of the compounds identified from this study with a low binding energy were proven as inhibitors of AICAR transformylase by *in vitro* experi-

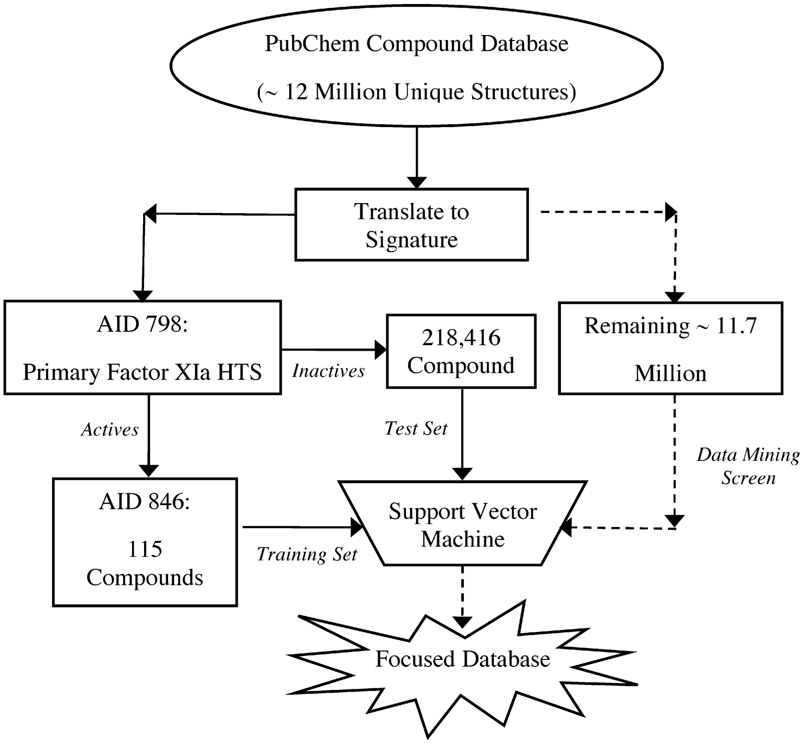


Fig. 2. A simple flowsheet describing the algorithm used to obtain the SVM and the focused database as a result of screening PubChem.

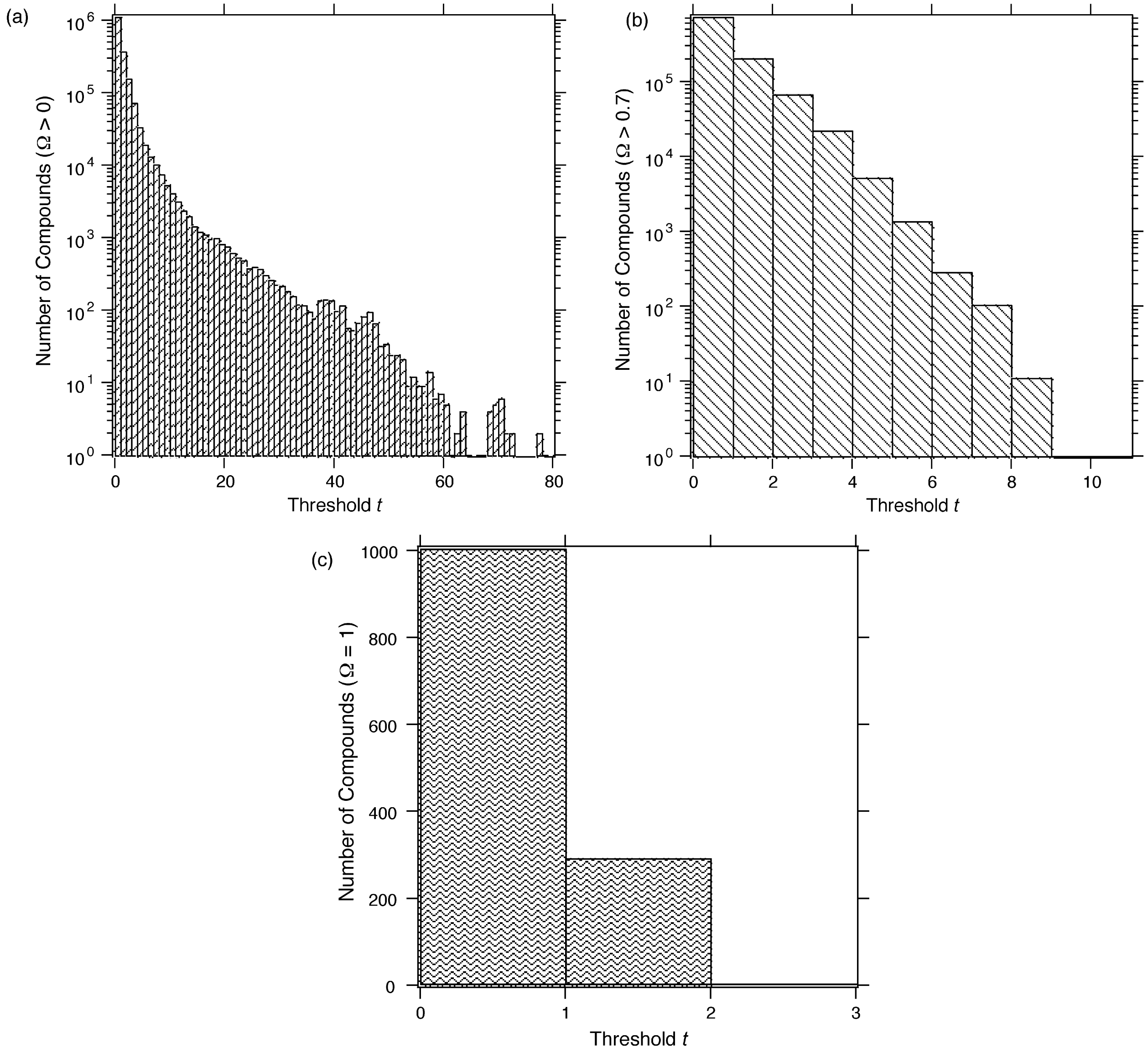


Fig. 3. Distribution of compounds predicted as active from the PCCompound database for the overlap metrics of (a) $\Omega > 0$, (b) $\Omega > 0.7$, and (c) $\Omega = 1$.

ments. As previously mentioned, predictions from the lower right-hand side of Table 3 are considered more reliable candidates to inhibit factor XIa. Because of this, and computational considerations (re: each run averaged one day), the 296 compounds where $\Omega = 1$ with threshold $t > 1$ were selected for docking.

The results of this portion of the docking study were very consistent with the predictions of activity from the SVM model for these potential new inhibitors. In particular, the minimum binding energies for this subset ranged from −5.48 to −9.84 kcal/mol, which is very close to the training set compound range. Accordingly, these compounds then would be considered possible inhibitors on the basis of this binding energy (in addition to the SVM screen). Such evidence from an independent test of these potential inhibitors for factor XIa provides additional confidence in the SVM model. Note that the binding energy for 12 of the 296 compounds is provided in Table 4 while a docked image for CID 7643488 is presented in Fig. 4.

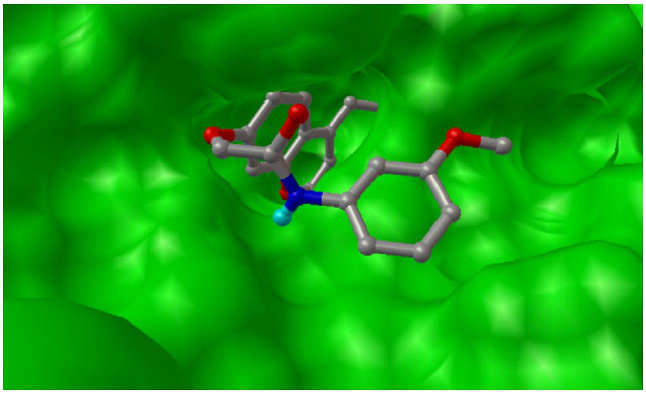


Fig. 4. Solvent-accessible molecular surface image of CID 7643488 docked (E_{Binding} = −9.2 kcal/mol) to the factor XIa active site.

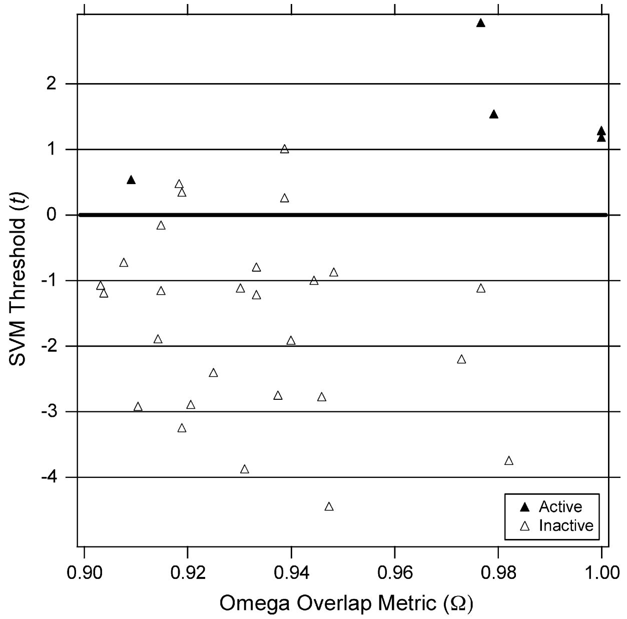


Fig. 5. A test set of 33 compounds was obtained from additional screening in AID 846. The bold horizontal line represents the default SVM threshold of zero, and all active compounds were correctly classified here.

3.4. Additional test set

While preparing the manuscript, additional compounds were screened in AID 846, and had not previously been included in the SVM training set. Instead of retraining the SVM with these new data, we utilized them as an independent test set to further evaluate our claim of confidence for predictions in the right side of Table 3. From this new data, a total of 33 compounds were available with an omega value greater than 0.90, where six were active ($IC_{50} < 5.0\ \mu M$) and 27 inactive ($IC_{50} > 5.0\ \mu M$). The SVM decision function value for this test set as a function of the omega overlap metric is provided in Fig. 5. Note that only five active compounds are visible since two observations had very similar SVM values (1.28 and 1.29) with an omega value of 1.0 and, thus, their symbols overlap in Fig. 5. The bold horizontal grid line indicates the default SVM threshold of zero. At this point, all active compounds are correctly classified while 23 of 27 inactive compounds are correctly classified resulting in a specificity of 0.85. Classification correctness is easily determined from Fig. 5. For active compounds, all above the threshold are classified correct (true positive), while those below are not correct (false negative). The classification correctness for inactive compounds can be evaluated in a related manner, except now correct predictions (true negative) fall below the threshold, and any incorrect (false positive) are above the threshold.

In a similar trend to Table 2, moving the threshold from zero to one in Fig. 5 allows for perfect precision for of all compounds active in this range, with no false positives. This result provides evidence that classification confidence in our SVM to identify true positives of factor XIa is high when there is both a high overlap metric (here, >0.9) and a large SVM threshold value (here, >1). Such a result on this test set reflects positively on the potential activity of those compounds screened from PubChem using this SVM in the previous section (re: those compounds in the bottom right of Table 3).

4. Discussion

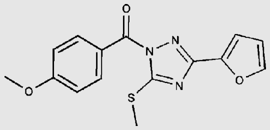
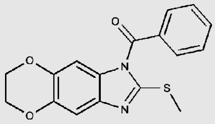
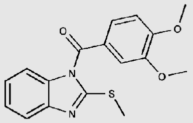
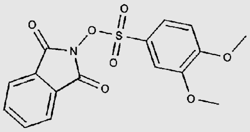
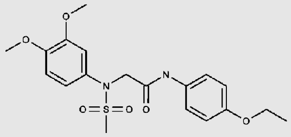
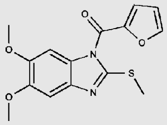
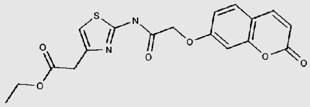
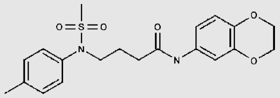
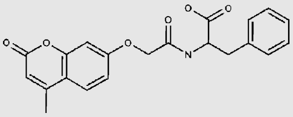
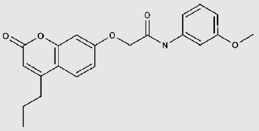
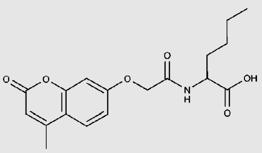
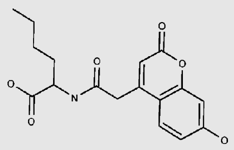
The ultimate goal of this work was to apply cheminformatic tools to utilize the HTS data available from PubChem, and suggest promising new candidates for additional consideration. Factor XIa was selected because it is an interesting therapeutic target, the 3-D structure is known, and a balanced training set was available. The HTS data from AID 846 was used to train the SVM classifier. Overfitting is always a concern with any statistical model, including SVM. When all Signatures available were included for SVM training, there was clearly overfitting since the 10-fold cross-validation accuracy was approximately that of a random classifier. Applying our improved version of recursive cluster elimination for feature selection significantly improved the SVM performance to 89% accuracy. Additional validation for the SVM was provided by two independent test sets. The first evaluated was an extremely large (218,416 compounds) set of known inactives from the primary HTS in AID 798. One might want to discount this as a test set since it does not contain any active compounds. However, this set provided a valuable function in that it showed the SVM was capable of avoiding false positive predictions, especially in the lower, right portion of Table 2 (re: where the threshold value and overlap metric are both high). The second test set, which recently became available because of additional screening in AID 846, was evaluated using the SVM which also provided perfect precision for the active compounds once the overlap metric and SVM threshold was large.

The trained and vetted SVM was used to virtually screen approximately 12 million compounds in PubChem that were not present in the factor XIa HTS. The ability to screen such a large set in a limited amount of time was an advantage of using SVM instead of docking alone. Most docking virtual screens are limited to search a smaller subset of compounds in a diversity database. The potential for extrapolation was a concern, and addressed via the overlap metric to provide a measure for common Signatures with the training set. All hits from the virtual screen were categorized by the overlap metric, and a direct comparison with the test sets was possible. From this analysis, it was concluded that predictions are likely more reliable toward the bottom right of Table 3 and thus, we direct focus here for further study. Additional confidence on the SVM results was provided by independent docking studies that were performed. As shown in Table 4, an interesting outcome was the diversity of potentially new inhibitors found. The compounds identified from this work form a focused database worthy of further study that could provide additional insights into factor XIa inhibition.

Table 3
Screening PubChem compounds for new factor XIa inhibitors with SVM. Predicted percent active, per the listed overlap metric, is provided in square brackets

t	$\Omega > 0$ (*n* = 11,946,913)	$\Omega > 0.6$ (*n* = 10,620,294)	$\Omega > 0.7$ (*n* = 9,020,826)	$\Omega > 0.8$ (*n* = 5,594,590)	$\Omega > 0.9$ (*n* = 1,378,787)	$\Omega = 1.0$ (*n* = 31,267)
>0	1,828,891 [15.3]	1,345,179 [12.7]	1,028,078 [11.4]	514,022 [9.2]	91,426 [6.6]	1300 [4.1]
>1	715,208 [6.0]	424,227 [4.0]	300,495 [3.3]	136,455 [2.4]	23,810 [1.7]	296 [0.9]
>2	343,052 [2.9]	149,199 [1.4]	96,063 [1.1]	37,470 [0.6]	4,899 [0.4]	4 [0.01]

Table 4
A sample of one dozen compounds from the $\Omega = 1$ set

CID	Structure	TC	SVM	E_{Binding} (kcal/mol)
3658123		0.94	1.29	−7.64
4426757		0.88	1.81	−7.13
977731		0.80	1.76	−7.45
2133598		0.75	1.61	−8.07
1098141		0.69	1.45	−7.28
10448578		0.62	2.17	−5.81
16418311		0.50	1.90	−8.23
6499012		0.48	1.60	−8.26
1184659		0.42	1.55	−9.20
7643488		0.42	1.06	−9.20
1556303		0.38	1.63	−7.62
6104501		0.37	1.85	−6.98

Reported are the compound ID number from PubChem, 2D-structure, maximum Tanimoto Coefficient (with AID 846), magnitude of the decision function (SVM), and binding energy (kcal/mol) from AutoDock.

5. Conclusion

In conclusion, this work presents a technique which combines a wrapper method which removes clusters for feature selection within an SVM with that of an overlap metric to identify potential active compounds against factor XIa available within PubChem. It is important to note that the technique in this work is not specific for factor XIa, and could be applied to other bioassays in PubChem to broaden the diversity of active compounds. PubChem provided significant resources that were conveniently available to serve this study. As additional bioassays and compounds are deposited in PubChem, the potential for data mining will continue to increase.

6. Supporting information

A list of CIDs for threshold t greater than zero from Table 3 is provided for all Ω values greater than 0.9. The first column in each file is the CID number, and the second column is the corresponding SVM decision function sorted by decreasing magnitude.

Acknowledgements

This work was funded by the Department of Energy's 2004 Presidential Early Career Scientist and Engineer Award. Computational work on this project was performed at the Computer Aided Engineering Laboratory on the Tennessee Technological University campus.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmgm.2008.08.004.

References

[1] C.P. Austin, L.S. Brady, T.R. Insel, F.S. Collins, NIH molecular libraries initiative, Science 306 (2004) 1138–1139.
[2] E. Zerhouni, Medicine, The NIH roadmap, Science 302 (2003) 63–72.
[3] Molecular Libraries Screening Centers Network. http://mli.nih.gov/mlscn/ (Accessed January 7, 2008).
[4] D.L. Wheeler, T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, R. Edgar, S. Federhen, L.Y. Geer, Y. Kapustin, O. Khovayko, D. Landsman, D.J. Lipman, T.L. Madden, D.R. Maglott, J. Ostell, V. Miller, K.D. Pruitt, G.D. Schuler, E. Sequeira, S.T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R.L. Tatusov, T.A. Tatusova, L. Wagner, E. Yaschenko, Database resources of the National Center for Biotechnology Information, Nucleic Acids Res. 35 (2007) D5–D12.
[5] T.I. Oprea, A. Tropsha, J.L. Faulon, M.D. Rintoul, Systems chemical biology, Nat. Chem. Biol. 3 (2007) 447–450.
[6] X.Q. Xie, J.Z. Chen, Data mining a small molecule drug screening representative subset from NIH PubChem, J. Chem. Inf. Model. 48 (2008) 465–475.
[7] Y. Zhou, B. Zhou, K. Chen, S.F. Yan, F.J. King, S. Jiang, E.A. Winzeler, Large-scale annotation of small-molecule libraries using public databases, J. Chem. Inf. Model. 47 (2007) 1386–1394.
[8] G.R. Rosania, G. Crippen, P. Woolf, D. States, K. Shedden, A cheminformatic toolkit for mining biomedical knowledge, Pharm. Res. 24 (2007) 1791–1802.
[9] S. Ingsriswang, E. Pacharawongsakda, sMOL Explorer: an open source, web-enabled database and exploration tool for Small MOLecules datasets, Bioinformatics 23 (2007) 2498–2500.
[10] F. Fontaine, E. Bolton, Y. Borodina, S.H. Bryant, Fast 3D shape screening of large chemical databases through alignment-recycling, Chem. Cent. J. 1 (2007) 12.
[11] Q. Li, F.S. Jorgensen, T. Oprea, S. Brunak, O. Taboureau, hERG classification model based on a combination of support vector machine method and GRIND descriptors, Mol. Pharm. 5 (2008) 117–127.
[12] V. Vapnik, Statistical Learning Theory, Wiley Interscience, New York, 1998.
[13] D. Plewczynski, M. von Grotthuss, S.A. Spieser, L. Rychlewski, L.S. Wyrwicz, K. Ginalski, U. Koch, Target specific compound identification using a support vector machine, Comb. Chem. High Throughput Screen 10 (2007) 189–196.

[14] M. Glick, J.L. Jenkins, J.H. Nettles, H. Hitchings, J.W. Davies, Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and Laplacian-modified naive bayesian classifiers, J. Chem. Inf. Model. 46 (2006) 193–200.

[15] R.N. Jorissen, M.K. Gilson, Virtual screening of molecular databases using a support vector machine, J. Chem. Inf. Model. 45 (2005) 549–561.

[16] E. Byvatov, G. Schneider, SVM-based feature selection for characterization of focused compound collections, J. Chem. Inf. Comput. Sci. 44 (2004) 993–999.

[17] M.K. Warmuth, J. Liao, G. Ratsch, M. Mathieson, S. Putta, C. Lemmen, Active learning with support vector machines in the drug discovery process, J. Chem. Inf. Comput. Sci. 43 (2003) 667–673.

[18] V.V. Zernov, K.V. Balakin, A.A. Ivaschenko, N.P. Savchuk, I.V. Pletnev, Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions, J. Chem. Inf. Comput. Sci. 43 (2003) 2048–2056.

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

[20] T. Eitrich, A. Kless, C. Druska, W. Meyer, J. Grotendorst, Classification of highly unbalanced CYP450 data of drugs using cost sensitive machine learning techniques, J. Chem. Inf. Model. 47 (2007) 92–103.

[21] C.W. Yap, Y.Z. Chen, Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines, J. Chem. Inf. Model. 45 (2005) 982–992.

[22] J.M. Kriegl, T. Arnhold, B. Beck, T. Fox, A support vector machine approach to classify human cytochrome P450 3A4 inhibitors, J. Comput. Aided Mol. Des. 19 (2005) 189–201.

[23] M. Zheng, Z. Liu, C. Xue, W. Zhu, K. Chen, X. Luo, H. Jiang, Mutagenic probability estimation of chemical compounds by a novel molecular electrophilicity vector and support vector machine, Bioinformatics 22 (2006) 2099–2106.

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

[25] H. Li, C.W. Yap, C.Y. Ung, Y. Xue, Z.W. Cao, Y.Z. Chen, Effect of selection of molecular descriptors on the prediction of blood–brain barrier penetrating and nonpenetrating agents by statistical learning methods, J. Chem. Inf. Model. 45 (2005) 1376–1384.

[26] J.L. Faulon, Stochastic generator of chemical structure. 1: Application to the structure elucidation of large molecules, J. Chem. Inf. Comput. Sci. 34 (1994) 1204–1218.

[27] J.L. Faulon, M.J. Collins, R.D. Carr, The signature molecular descriptor. 4: Canonizing molecules using extended valence sequences, J. Chem. Inf. Comput. Sci. 44 (2004) 427–436.

[28] S. Martin, D. Roe, J.L. Faulon, Predicting protein–protein interactions using signature products, Bioinformatics 21 (2005) 218–226.

[29] D.W. Chung, K. Fujikawa, B.A. McMullen, E.W. Davie, Human plasma prekallikrein, a zymogen to a serine protease that contains four tandem repeats, Biochemistry 25 (1986) 2410–2417.

[30] K. Naito, K. Fujikawa, Activation of human blood coagulation factor XI independent of factor XII. Factor XI is activated by thrombin and factor XIa in the presence of negatively charged surfaces, J. Biol. Chem. 266 (1991) 7353–7358.

[31] A. Gruber, S.R. Hanson, Potential new targets for antithrombotic therapy, Curr. Pharm. Des. 9 (2003) 2367–2374.

[32] Factor XIa, 1536 HTS. http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=798 (Accessed January 14, 2008).

[33] Factor XIa 1536 HTS Dose Response Confirmation. http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=846 (Accessed January 14, 2008).

[34] L. Jin, P. Pandey, R.E. Babine, J.C. Gorga, K.J. Seidl, E. Gelfand, D.T. Weaver, S.S. Abdel-Meguid, J.E. Strickler, Crystal structures of the FXIa catalytic domain in complex with ecotin mutants reveal substrate-like interactions, J. Biol. Chem. 280 (2005) 4704–4712.

[35] H. Deng, T.D. Bannister, L. Jin, R.E. Babine, J. Quinn, P. Nagafuji, C.A. Celatka, J. Lin, T.I. Lazarova, M.J. Rynkiewicz, F. Bibbins, P. Pandey, J. Gorga, H.V. Meyers, S.S. Abdel-Meguid, J.E. Strickler, Synthesis, SAR exploration, and X-ray crystal structures of factor XIa inhibitors containing an alpha-ketothiazole arginine, Bioorg. Med. Chem. Lett. 16 (2006) 3049–3054.

[36] T.I. Lazarova, L. Jin, M. Rynkiewicz, J.C. Gorga, F. Bibbins, H.V. Meyers, R. Babine, J. Strickler, Synthesis and in vitro biological evaluation of aryl boronic acids as potential inhibitors of factor XIa, Bioorg. Med. Chem. Lett. 16 (2006) 5022–5027.

[37] J. Lin, H. Deng, L. Jin, P. Pandey, J. Quinn, S. Cantin, M.J. Rynkiewicz, J.C. Gorga, F. Bibbins, C.A. Celatka, P. Nagafuji, T.D. Bannister, H.V. Meyers, R.E. Babine, N.J. Hayward, D. Weaver, H. Benjamin, F. Stassen, S.S. Abdel-Meguid, J.E. Strickler, Design, synthesis, and biological evaluation of peptidomimetic inhibitors of factor XIa as novel anticoagulants, J. Med. Chem. 49 (2006) 7781–7791.

[38] C.J. Churchwell, M.D. Rintoul, S. Martin, D.P. Visco Jr., A. Kotu, R.S. Larson, L.O. Sillerud, D.C. Brown, J.L. Faulon, The signature molecular descriptor. 3: Inverse-quantitative structure-activity relationship of ICAM-1 inhibitory peptides, J. Mol. Graph. Model. 22 (2004) 263–273.

[39] J.L. Faulon, C.J. Churchwell, D.P. Visco Jr., The signature molecular descriptor. 2: Enumerating molecules from their extended valence sequences, J. Chem. Inf. Comput. Sci. 43 (2003) 721–734.

[40] J.L. Faulon, D.P. Visco Jr., R.S. Pophale, The signature molecular descriptor. 1: Using extended valence sequences in QSAR and QSPR studies, J. Chem. Inf. Comput. Sci. 43 (2003) 707–720.

[41] D.P. Visco Jr., R.S. Pophale, M.D. Rintoul, J.L. Faulon, Developing a methodology for an inverse quantitative structure-activity relationship using the signature molecular descriptor, J. Mol. Graph. Model. 20 (2002) 429–438.

[42] L.H. Hall, MOLCONN-Z, Hall Associates Consulting, Quincy, MA, 1991.

[43] J.C. Burgess, A tutorial on support vector machines for pattern recognition, Data. Min. Knowl. Disc. 2 (1998) 121–167.

[44] T. Joachims, Making large-scale SVM learning practical, in: Advances in Kernel Methods—Support Vector Learning, MIT Press, 1999.

[45] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.

[46] T.S. Furey, N. Cristianini, N. Duffy, D.W. Bednarski, M. Schummer, D. Haussler, Support vector machine classification and validation of cancer tissue samples using microarray expression data, Bioinformatics 16 (2000) 906–914.

[47] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1997) 273–324.

[48] M. Yousef, S. Jung, L.C. Showe, M.K. Showe, Recursive cluster elimination (RCE) for classification and feature selection from gene expression data, BMC Bioinform. 8 (2007) 144.

[49] D. Wishart, Clusteran: A Class Act, Clusteran Limited, Edinburgh, 2003.

[50] R. Barandela, J.S. Sanchez, V. Garcia, E. Rangel, Strategies for learning in class imbalance problems, Pattern Recog. 36 (2003) 849–851.

[51] D.J. Diller, D.W. Hobbs, Deriving knowledge through data mining high-throughput screening data, J. Med. Chem. 47 (2004) 6373–6383.

[52] Marvin Beans 4.1.5, ChemAxon Ltd., Budapest, 2007.

[53] J. Gasteiger, T. Engel, Chemoinformatics, Wiley-VCH, Weinheim, 2003.

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

[55] C. Li, L. Xu, D.W. Wolan, I.A. Wilson, A.J. Olson, Virtual screening of human 5-aminoimidazole-4-carboxamide ribonucleotide transformylase against the NCI diversity set by use of AutoDock to identify novel nonfolate inhibitors, J. Med. Chem. 47 (2004) 6681–6690.