Pharmaceutical Research, Vol. 26, No. 9, September 2009 (© 2009)
DOI: 10.1007/s11095-009-9937-8

Research Paper

Predicting Inhibitors of Acetylcholinesterase by Regression and Classification Machine Learning Approaches with Combinations of Molecular Descriptors

Dmitriy Chekmarev,[1] Vladyslav Kholodovych,[1] Sandhya Kortagere,[1,4] William J. Welsh,[1,5] and Sean Ekins[1,2,3,5]

Received February 15, 2009; accepted July 2, 2009; published online July 15, 2009

Purpose. Acetylcholinesterase (AChE) is both a therapeutic target for Alzheimer's disease and a target for organophosphorus, carbamates and chemical warfare agents. Prediction of the likelihood of compounds interacting with this enzyme is therefore important from both therapeutic and toxicological perspectives.

Materials and Methods. Support vector machine classification and regression models with molecular descriptors derived from Shape Signatures and the Molecular Operating Environment (MOE) application software were built and tested using a set of piperidine AChE inhibitors (N=110).

Results. The combination of the alignment free Shape Signatures and 2D MOE descriptors with the Support Vector Regression method outperforms the models based solely on 2D and internal 3D (i3D) MOE descriptors, and is comparable with the best previously reported PLS model based on CoMFA molecular descriptors ($r^2_{test,SVR} = 0.48$ *vs.* $r^2_{test,PLS} = 0.47$ from Sutherland *et al.* J Med Chem 47:5541–5554, 2004). Support Vector Classification algorithms proved superior to a classifier based on scores from the molecular docking program GOLD, with the overall prediction accuracies being $Q_{SVC(10CV)}$=74% and $Q_{SVC(LNO)}$=67% *vs.* Q_{GOLD}=56%.

Conclusions. These new machine learning models with combined descriptor schemes may find utility for predicting novel AChE inhibitors.

KEY WORDS: acetylcholinesterase; docking; machine learning; molecular operating environment; quantitative structure activity relationship; shape signatures descriptors; support vector classification; support vector machine; support vector regression.

INTRODUCTION

Acetylcholinesterase (AChE) has long been considered a therapeutic target in symptomatic treatment of Alzheimer's disease to treat cognitive deficiency. The inhibitors of this enzyme work by reversibly blocking binding of the substrate to AChE or by hydrolytic inactivation of AChE high affinity site, thus effectively increasing the concentration of the neutrotransmitter acetylcholine (ACh) in nerve endings. A number of FDA approved AChE drugs include tacrine, rivastigmine, donepezil and galantamine. In addition, organophosphorus (OP) compounds, which have found uses ranging from insecticides to chemical warfare agents (CWAs), exert their toxic effects by reacting irreversibly with a catalytic serine to inhibit AChE. This in turn results in overproduction of acetylcholine at cholinergic synapses and overstimulation of muscarinic and nicotine receptors (1). Various crystal structures have also facilitated computational design of these and additional ligands at both the catalytic and peripheral binding sites (2). For example donepezil is positioned to occupy an anionic subsite in the active site gorge by engaging in π-π stacking and cation-π interactions (3). Synthetic efforts employing medicinal chemistry have been a fertile source of multiple classes of AChE inhibitors (2), which in turn have been employed to build quantitative structure activity relationship (QSAR) models (4). Using Comparative Molecular Field Analysis (CoMFA) (5), one study constructed a model using 57 *N*-benzylpiperidines in the training set (cross validated r^2 values >0.6) and 20 molecules in the test set (r^2 >0.8) (6). A later study that used a protein receptor-based alignment, docking *N*-benzylpiperidines and related compounds in the mouse AChE structure, yielded a comparable CoMFA model with cross validated r^2 values from 0.6 to >0.7 (7). A series of eight organophosphorus AChE inhibitors were used to generate a Catalyst

Electronic supplementary material The online version of this article (doi:10.1007/s11095-009-9937-8) contains supplementary material, which is available to authorized users.

[1] Department of Pharmacology and Environmental Bioinformatics & Computational Toxicology Center (ebCTC), University of Medicine & Dentistry of New Jersey, Robert Wood Johnson Medical School, 675 Hoes Lane, Piscataway, New Jersey 08854, USA.

[2] Collaborations in Chemistry, 601 Runnymede Avenue, Jenkintown, Pennsylvania 19046, USA.

[3] Department of Pharmaceutical Sciences, University of Maryland, College Park, Maryland 21201, USA.

[4] Department of Microbiology and Immunology, Drexel University College of Medicine, Philadelphia, Pennsylvania 19129, USA.

[5] To whom correspondence should be addressed. (e-mail: welshwj@umdnj.edu; ekinssean@yahoo.com)

pharmacophore consisting of a ring aromatic feature, two hydrophobic features and a hydrogen bond acceptor, with a training correlation r^2=0.994. This model was used to search a database of >30,000 compounds to identify further compounds to test (8). Another study applied a large array of QSAR methods on >100 AChE inhibitors (and seven other sets of compounds) by splitting the dataset into training and test sets to compare their results (Supplemental Table 1). CoMFA appeared to perform the best (9). More recently Bayesian-regularized genetic neural networks were used with Dragon descriptors to provide a QSAR for 60 huprines (10). Docking scores obtained with FlexX and FlexiDock for a series of 19 bis-tacrines were used as supplement CoMFA descriptors to yield a model with cross validated r^2=0.71. The steric and electrostatic fields were also superimposed in the crystal structure to assist inhibitor design (11). A further study on a dataset of 80 AChE inhibitors split into training (68) and testing (12) sets (encompassing several sets of published tacrines, huprines and other compounds) using a total of 133 descriptors with stepwise multiple linear regression alone or combined with simulated annealing and genetic algorithms, resulted in test set correlations between r^2=0.73–0.84. Several of the descriptors used in the final models were Kier shape descriptors (12). Finally, in a recent study, Manchester and Czermiński demonstrated that machine learning methods, such as Random Forest and Support Vector Regression, combined with alignment-based molecular descriptors, can efficiently model AChE inhibition activities of certain organic molecules (13). Based on these and other previously published studies, QSAR methods are important tools for pharmaceutical drug design and for screening of potential chemical warfare agents.

The goal of the current study was to examine the utility of the alignment independent Shape Signatures (SS) and MOE descriptors in the context of quantitative (regression) and qualitative (classification) predictive models for molecular interaction with the AChE receptor. Previously, these sets of molecular descriptors have been successfully applied to cardiotoxicity-related target proteins and blood-brain barrier data with machine learning classification method (14,15). In these studies, it was found that 2DSS descriptors slightly outperformed 1DSS with the SVM algorithm and that SVM models based on Shape Signatures also performed slightly better than those developed with the MOE descriptors for the same datasets (14,15). Here, we demonstrate that a similar collection of shape-based and alignment-free molecular descriptors can be used to build predictive models for AChE inhibition. When paired with the Support Vector Machine regression (SVR) algorithm, the resulting QSAR models performed competitively with more compute-intensive alignment-based methods such as CoMFA/PLS. Therefore, novel shape-based, alignment-independent molecular descriptors are a practical alternative to CoMFA and other alignment-based approaches for future QSAR modeling. In addition, we have reported a rigorous classification procedure which can be used to categorize the drug-like molecules as either strong or weak AChE inhibitors. This new approach was compared with classification based on docking scores obtained from GOLD (16).

MATERIALS AND METHODS

Datasets

The published dataset from Sutherland *et al.* (9) of 110 piperidine derivatives split into training (73) and test (37) sets was used for this study. The list of original references and the associated chemical structures can be found in Sutherland *et al.* (9). The compounds in the test set were carefully selected from the original pool of structures, and both the training and test sets are structurally diverse, covering the entire range of the measured activity data. The training and test sets were utilized with SVR and PLS regression studies to allow a quantitative comparison with previously reported PLS regression models (9). For classification analysis, the entire set of 110 structures was subjected to statistical analysis to assess model quality as described below. Using SMILES strings taken from the original papers as input, for each compound a single default low energy conformation was generated by CORINA (Molecular Networks GmbH, Erlangen, German, http://www.mol-net.de) and assigned partial atomic charges according to the Gasteiger-Marsili scheme (17). The resulting molecular conformations, one per compound, were then used to generate Shape Signatures histograms and to compute MOE molecular descriptors.

Molecular Descriptors for Regression and Classification

Two classes of molecular descriptors were considered in this study. The first group comprised 184 2D and 29 internal

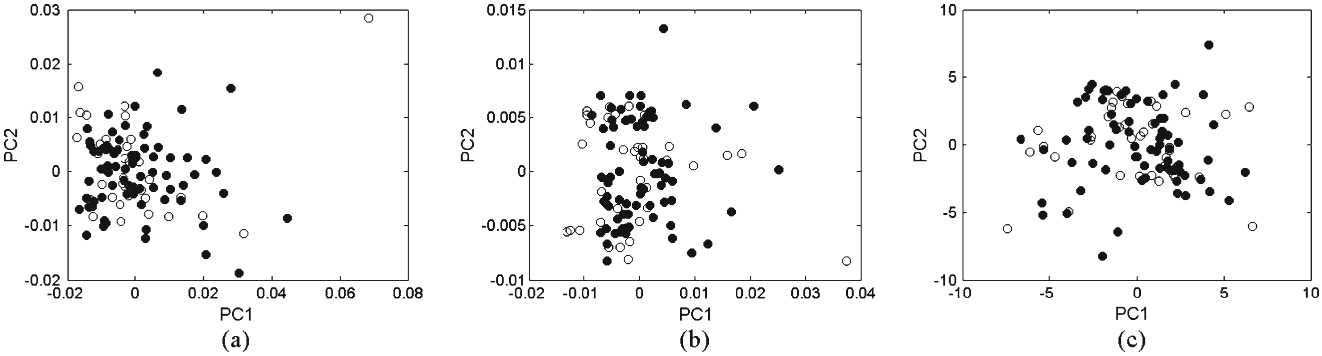

Fig. 1. PCA analysis performed in the space of 1DSS (**a**), 2DSS (**b**) and 1DSS + MOE2D (**c**) molecular descriptors. *Filled circles*: molecules from the training set. *Open circles*: molecules from the test set.

3D (i3D) molecular descriptors calculated using the Molecular Operating Environment (MOE, Chemical Computing Group, Montreal, Canada) modeling program. The latter set comprised descriptors that depend on the molecular conformation and includes a number of quantum chemical descriptors, such as the molecular dipole moments, heats of formation, ionization potentials, HOMO and LUMO energies calculated using the AM1 semi-empirical method. It should be noted that all 213 MOE molecular descriptors used in this study are alignment independent, i.e. they do not require the molecules to be orientated in any particular way prior to descriptor generation (such as in CoMFA, which requires molecular alignment).

The second group of descriptors is the shape-based descriptors derived from the Shape Signatures method (1D shape signature: 1DSS; 2D shape signature: 2DSS) (18), a method that has been documented extensively in the literature (14,15,18–20). Shape Signatures employs a ray-tracing algorithm to explore the volume enclosed by the solvent-accessible surface of a molecule, and converts this information into simple numeric representations (i.e., signatures) that encode molecular shape and polarity. The degree of similarity between a pair of molecules can be assessed by comparing their 1D signatures (shape only) or 2D signatures (shape and surface charge distribution). This process is fast and efficient, and it eliminates tedious and subjective atom-based alignment of the molecules required in many traditional molecular modeling approaches. Shape Signatures is amenable to numerous molecular modeling and cheminformatics applications, including drug discovery, virtual database screening, and predictive toxicology (14,15,18–23). Similar to previous applications (14,15,20) for each compound in this study, the heights of the bins of the associated 1DSS (shape only) and 2DSS (shape and polarity) Shape Signature histograms represent distinct molecular descriptors which are inherently three-dimensional and are also alignment independent.

Finally, we have examined a number of mixed descriptor schemes where we combined molecular descriptors from Shape Signatures with those computed with MOE. In particular, below we report the results of regression and classification analyses for the following mixed descriptor libraries: 1DSS + MOE2D, 1DSS + MOE2Di3D and 2DSS + MOE2Di3D.

Data Preparation for Regression and Classification

The large collections of molecular descriptors described above were filtered with the unsupervised forward selection (UFS) procedure of Livingstone and co-workers (24) prior to running SVR, PLS or SVC to remove redundant and correlated descriptors. The UFS program has proven valuable for data preparation for subsequent QSAR (24) and classification analysis (14,15,20). The final compositions of all descriptor selection schemes considered in this study are as follows: 22 descriptors for 1DSS, 53 descriptors for 2DSS, 42 descriptors for MOE2D, 45 descriptors for MOE2Di3D (91% 2D MOE and 9% i3D MOE), 54 descriptors for 1DSS + MOE2D (69% 2D MOE and 31% 1DSS), 55 descriptors for 1DSS + MOE2Di3D (66% 2D MOE, 7% i3D MOE and 27% 1DSS) and 60 descriptors for 2DSS + MOE2Di3D (57% 2D MOE, 3% i3D MOE and 40% 2DSS). We note that the Shape Signatures descriptors make up a large fraction of descriptors in the mixed descriptor libraries, although lower than the 2D MOE descriptors.

The utility of predictive models may suffer due to the unique regions of chemical space (i.e., applicability domain) occupied by structures from the training and test sets. As we have introduced new sets of molecular descriptors in this work, it was therefore important to examine the chemical space covered by these new molecular descriptors. We conducted principal component analysis (PCA) with 1DSS, 2DSS and 1DSS + MOE2D molecular descriptors. The first few PCs, which are certain linear combinations of the input descriptors, define the directions along which the data points (molecules) are maximally spread. If the structures from the training and test sets do indeed occupy remote regions of the chemical space, which would generally undermine the quality of predictions, this should become visible in the first PCs as these will reflect maximal separations between the data points. Fig. 1a–c display the positions of the molecules from the training and test sets on the PC1-PC2 plane for 1DSS (variance explained: PC1 = 60% and PC2 = 17%), 2DSS

Table I. PLS Regression Analyses of Ache Compounds from Sutherland *et al.* (9). The Training Set Contains 73 Molecules and the Test Set Contains 37 Molecules

Molecular descriptors[a]	q^2_{PLS} [b]	s_{PLS}[b]	Number of PLS components[b]	r^2_{train} [c]	s_{train}[c]	r^2_{test} [d]	s_{test}[d]
1DSS (22)	−0.56	1.48	1	–	–	–	–
2DSS (53)	−0.09	1.27	1	–	–	–	–
MOE2D(42)	0.31	1.02	2	0.55	0.83	0.34	1.09
MOE2Di3D (45)	0.42	0.93	2	0.64	0.75	0.42	1.03
1DSS + MOE2D (54)	0.28	1.04	3	0.64	0.75	0.43	1.03
1DSS + MOE2Di3D (55)	0.31	1.01	3	0.67	0.72	0.44	1.02
2DSS + MOE2Di3D (60)	0.22	1.08	3	0.64	0.75	0.38	1.08

[a] Numbers in parenthesis reflect the number of independent molecular descriptors selected by the UFS (unsupervised forward selection) algorithm.

[b] These three columns list q^2_{PLS} , s_{PLS} (analogous to s_{PRESS}) and the model complexity, i.e., the number of the PLS components, for the best PLS model determined from 10-fold cross validations performed on the training set. The best PLS model minimizes s_{PLS} (maximizes q^2_{PLS}).

[c] These two columns list r^2_{train} and s_{train} for the best PLS model determined from 10-fold cross validations.

[d] These two columns list r^2_{test} and s_{test} for the best PLS model determined from 10-fold cross validations.

(variance explained: PC1 = 47% and PC2 = 17%) and 1DSS + MOE2D (variance explained: PC1 = 15% and PC2 = 15%). For the latter case, similar behavior was observed when we examined the locations of the data points in the coordinate systems of the other principal components (PC1 *vs*. PC3, PC2 *vs*. PC3, etc.). The test and training sets would appear well distributed in the molecular descriptor space suggesting that when a predictive model is developed with the training set, it can be applied to predict the test set with little concern about uncertainty due to exceeding the applicability domain.

Regression by Partial Least Squares (PLS)

Partial Least Squares (25) is a popular linear regression tool that can deal with large input sets of correlated descriptors and which has been routinely used in chemometrics and cheminformatics. The method includes features of both principal component analysis and multiple linear regression, generating combinations of descriptors with large variances and reduced intercomponent correlations which simultaneously show large covariance with the endpoint (activity) data.

In this study, for each descriptor library the data from the training set was used to identify a single best performing PLS model, similar to the protocol outlined in (9), in which the best PLS model was the one which recorded the lowest root mean-squared error in the 10-fold cross correlation (s_{10CV}) experiment across the entire training set. Following this previous study (9), the s_{10CV} values were modified before final selection in order to account for the complexity of the model reflected in the number of the PLS components. Once generated, this model was applied to predict the data from both the training and test sets. The quality of such predictions for each descriptor selection scheme was monitored by computing the root mean squared error (s_{train} and s_{test}) and the squared correlation coefficient $\left(r^2_{train}\ and r^2_{test}\right)$ between the observed and predicted data. The output PLS models contained typically 1 to 3 PLS components. The PLS as well as PCA analyses reported earlier were carried out using the routines from the Statistics Toolbox of MATLAB (Version 7.6, The MathWorks, Inc, Natick, MA).

Support Vector Classification and Support Vector Regression

The SVM method (26,27) is a powerful machine learning technique that has been used widely to tackle complex classification and regression problems (13,26–35). SVM is based on the structural risk minimization principle which seeks to minimize the upper bound on the expected risk (26,27), and thus is believed to display better generalization properties than methods derived from empirical risk minimization. The SVM binary classification procedure is capable of separating the representatives from two classes even when they are linearly inseparable in the space of input descriptors. This is accomplished by projecting the data into a higher dimensional feature space where linear separation is frequently possible. SVM uses a special type of function, called the kernel function, which represents the interactions between the data points in high dimensional space. Following our previous studies (14,15,20), for classification we have used the freely available program LIBSVM (C-SVM) (36) with the Gaussian radial basis function kernel, whose parameter γ along with the SVM penalty term C, were determined in each case through a simple grid search procedure by 10-fold cross validation.

For classification, we combined data from the training and test sets into a single dataset of 110 molecules. For each choice of descriptor library, the UFS procedure was carried out for the entire dataset. The composition and the numbers of the retained descriptors were found to change slightly from those reported above. The dividing boundary was set at IC_{50}=150 nM, which resulted in 56 strong and 54 weak AChE inhibitors. Calculations for other representative boundaries 100, 250 and 500 nM are provided in the Supplementary Material (Supplemental Tables 2, 3, 4). The quality of all classification models was evaluated by considering the following set of statistical indicators: sensitivity (SE), specificity (SP), overall prediction accuracy (Q) and Matthews correlation coefficient (C) (33,37). These quantities are defined in terms of the numbers of true positives (TP; true strong AChE inhibitors), false positives (FP; weak AChE inhibitors classified as strong), true negatives (TN; true weak AChE inhibitors), and false negatives (FN; strong AChE inhibitors classified as weak). In these notations, the total number of real experimentally documented activators is given

Table II. ε-SVR (support vector regression) Analysis of AChE Compounds from Sutherland *et al.* (9). The Training Set Contains 73 Molecules and the Test Set Contains 37 Molecules

Molecular descriptors[a]	s_{10CV}[b]	r^2_{train}[c]	s_{train}[c]	r^2_{test}[d]	s_{test}[d]
1DSS (22)	0.99	1.00	1.03×10^{-3}	0.40	1.01
2DSS (53)	0.96	0.87	0.44	0.30	1.14
MOE2D (42)	1.03	0.93	0.36	0.36	1.04
MOE2Di3D (45)	0.97	0.96	0.24	0.36	1.05
1DSS + MOE2D (54)	0.92	0.99	0.12	0.48	0.97
1DSS + MOE2Di3D (55)	0.90	0.93	0.32	0.44	0.99
2DSS + MOE2Di3D (60)	0.94	0.90	0.38	0.41	1.03

[a] Numbers in parenthesis reflect the number of independent molecular descriptors selected by the UFS (unsupervised forward selection) algorithm.

[b] This column lists s_{10CV} for the best SVR model determined from the 10-fold cross validation performed on the training set. The best SVR model minimizes s_{10CV}.

[c] r^2_{train} and s_{train} for the best SVR model determined from 10-fold cross validations.

[d] r^2_{test} and s_{test} for the best SVR model determined from 10-fold cross validations.

Table III. SVM Classification of 110 AChE Compounds from Sutherland *et al.* (9). The Dividing Boundary was Set at IC_{50}=150 nM Resulting in 56 Strong and 54 Weak Inhibitors

		Leave-20%-out testing[b]			
Molecular descriptors	Q_{10CV}[a] (%)	⟨SE⟩ (%)	⟨SP⟩ (%)	⟨Q_{LNO}⟩ (%)	C
1DSS	69	52	67	59	0.193
2DSS	74	68	64	66	0.328
MOE2D	69	62	69	66	0.320
MOE2Di3D	73	61	70	65	0.313
1DSS + MOE2D	74	64	70	67	0.351
1DSS + MOE2Di3D	76	61	69	64	0.302
2DSS + MOE2i3D	76	66	68	67	0.343

[a] This column lists prediction accuracies Q_{10CV} estimated from 10-fold cross validations performed on the entire dataset.
[b] The tabulated values of ⟨SE⟩, ⟨SP⟩, ⟨Q_{LNO}⟩ and C were averaged over the results of 100 different hold-out test set experiments (see text for details).

by TP + FN, whereas a corresponding number of real non-activators is TN + FP. Sensitivity then expresses the prediction accuracy of a classification model with respect to AChE inhibitors: $SE = TP/(TP + FN)$; while specificity reflects the prediction accuracy for non-activators: $SP = TN/(TN + FP)$. The overall prediction accuracy is calculated as $Q = (TP + TN)/(TP + FP + TN + FN)$. Finally, we also report values of the Matthews correlation coefficient (33,37) described as $C = [TP \times TN - FP \times FN] \big/ [(TP + FN)(TP + FP)(TN + FP)(TN + FN)]^{1/2}$. For a perfect classifier with FP = FN = 0, one would have C = 1.0. For a random prediction, C≈0, and for a complete inversion (TP = TN = 0), C = −1.0.

Based on our previous work (14,15), two types of model validation were used to examine SVC models. The first group included models generated by 10-fold cross validation conducted on the entire dataset. The overall prediction accuracies for these models are denoted by Q_{10CV}. The models from the second group were averaged over a series of 100 independent leave-20%-out runs (overall accuracies ⟨Q_{LNO}⟩). The leave-20%-out tests were designed as follows: for each dataset, about 20% of the molecules were randomly picked to represent the hold-out test set and the rest of the data constituted the training set for this particular data division. The selection was carried out to approximately preserve the correct proportion of AChE activators and non-activators in both sets. Each SVM classification algorithm was then trained on the training set and applied to predict class attributes of the compounds in the test set. To obtain more reliable statistical estimates, the procedure was repeated 100 times, each time with a different composition of the test and training sets.

With inclusion of the distance-dependent loss function it is possible to apply the SVM to regression problems (13,29,34,35). The loss function tries to position the hyperplane in the feature space such that the data points lie maximally close to it. In this study we utilized the SVM algorithm with the ε-insensitive loss function (ε-SVR), where all the data lying outside the region of distance ε from the hyperplane in the feature space are penalized. We used the LIBSVM (ε-SVR) (36) program with two choices for the kernel function: the linear kernel and the Gaussian radial basis function kernel. In each case, the parameters of the model and the ε term were determined through a grid search by 10-fold cross validation. This protocol is similar to that described above for the PLS regression. For each descriptor library, after having identified the best ε-SVR model for the training set, we applied it to model the data in the test set. We also used the same set of statistical measures (s_{train}, r^2_{train} , s_{test}, and r^2_{test}) to judge the quality of the SVR predictions.

Table IV. Classification of 110 AChE Compounds from the Sutherland *et al.* Dataset Based on Docking with GOLD. The Dividing Boundary was Set at IC_{50}=150 nM Resulting in 56 Strong and 54 Weak Inhibitors

Method	SE (%)	SP (%)	Q (%)	C
GOLD scores	32	81	56	0.156

Docking and Scoring with GOLD

The docking program GOLD (Version 3.1) (16) was used to dock all 110 compounds to the binding site of the human isoform of AChE (PDB ID:1B41 (38)). For each ligand, 30 independent docking runs were performed in order to identify the top docking mode. The best ranking conformation for each ligand was chosen according to the corresponding GoldScore (16) of that conformation. This approach resembles the default scheme for a typical high-throughput docking experiment, and in this study we preferred it over more sophisticated procedures (39) which entail construction and validation of the scoring functions specifically designed for docking of the considered set of inhibitors to the AChE receptor. Classification was carried out based on the values of the GoldScore fitness function without any additional weighting factor. These scores varied from 50.94 to 82.19. The cutoff for classification was chosen as half the value of the best and the worst docking scores combined. According to this scheme, molecules with docking scores above this value were categorized as strong inhibitors and those below this boundary as weak inhibitors. The predicted compound classifications based on the GoldScore were compared with their experimental classification and the results of the SVM classification.

RESULTS

The published dataset from Sutherland *et al.* (9) of 110 piperidine derivatives was used for this study. First we compared our linear regression results using non-alignment-dependent descriptors with those for alignment-dependent descriptors used with PLS by Sutherland *et al.* (9) (Table I). The following three models were found to perform the best with respect to their ability to model AChE inhibition activity of the test set: PLS with 1DSS + MOE2D $\left(q^2_{PLS} = 0.28,\ r^2_{test} = 0.43\right)$, PLS with 1DSS + MOE2Di3D $\left(q^2_{PLS} = 0.31,\ r^2_{test} = 0.44\right)$, and PLS with MOE2Di3D $\left(q^2_{PLS} = 0.42,\ r^2_{test} = 0.42\right)$. All three alignment-free models performed comparably with the alignment-dependent CoMSIA $\left(q^2_{PLS} = 0.46,\ r^2_{test} = 0.44\right)$ and CoMFA $\left(q^2_{PLS} = 0.52,\ r^2_{test} = 0.47\right)$ models reported previously by Sutherland *et al.* (9) (Supplemental Table 1). However, we were unable to find reasonable PLS models for either 1DSS or 2DSS alone, based on their negative q^2_{PLS} obtained with the 10-fold cross validation.

In addition to PLS, we have evaluated the SVR algorithm with a linear kernel function. The best performing models were obtained with MOE2D and MOE2Di3D descriptors, which yielded $r^2_{test} = 0.39$ and $r^2_{test} = 0.35$, respectively. This performance is similar to that observed with the PLS models described above. Mixed descriptor schemes performed worse than with PLS resulting in $r^2_{test} \approx 0.23 - 0.26$, while 1DSS $\left(r^2_{test} = 0.11\right)$ and 2DSS $\left(r^2_{test} = 0.16\right)$ were able to generate only weakly positive correlations.

Further regression experiments utilized the non-linear SVR algorithm with the Gaussian radial basis function kernel. Here, data fitting is performed in a high dimensional feature space (Table II). Again the top performers were SVR models based on mixed descriptor schemes, i.e. SVR with 1DSS + MOE2D $\left(r^2_{test} = 0.48\right)$, SVR with 1DSS + MOE2Di3D $\left(r^2_{test} = 0.44\right)$ and SVR with 2DSS + MOE2Di3D $\left(r^2_{test} = 0.41\right)$. Interestingly, the SVR model using 1DSS alone $\left(r^2_{test} = 0.40\right)$ was comparable to these previously described models and shows a vast improvement over the linear kernel. The SVR models built on 1DSS and combined 1DSS and MOE2D descriptor schemes are therefore recommended as viable alternatives to CoMFA/PLS (9).

We have also constructed a series of SVC models to classify compounds according to the AChE inhibition activity. The dividing boundary was positioned at IC_{50}=150 nM, which yielded a balanced set of 56 strong and 54 weak inhibitors, and the results for various statistical tests were compared for different descriptor schemes (Table III). As before, the best performing method is the SVC coupled with 1DSS + MOE2D descriptors (Q_{10CV}=74%, $\langle Q_{LNO}\rangle$=67% and *C*=0.351), followed by the methods based on 2DSS + MOE2Di3D (Q_{10CV}=76%, $\langle Q_{LNO}\rangle$=67% and *C*=0.343) and 2DSS alone (Q_{10CV}=74%, $\langle Q_{LNO}\rangle$=66% and *C*=0.328). In contrast to the SVR regression models discussed above, SVC based on 1DSS alone performed the worst (Q_{10CV}=69%, $\langle Q_{LNO}\rangle$=59% and *C*=0.193). We have compared the SVM results with classification based on GOLD docking scores (Table IV). Classification of the actives and inactives based on raw docking scores at the 150 nM cutoff suggested that only 56% of them were predicted correctly. This observation can be attributed to a low positive correlation between the calculated GOLD scores and experimental activities pIC_{50} (*r*=0.27 and r^2=0.07). Thus, the comparison of SVC and classification based on GOLD revealed that SVC results were substantially more accurate.

Finally, we have compared the statistical regression models built using the top ranked 3D structures from GOLD (one conformation for each substance) with the models constructed using 3D molecular conformations generated by CORINA (one conformation for each substance), which ignores interactions with the receptor. Four sets of molecular descriptors, 1DSS, 2DSS, MOE2Di3D and 1DSS + MOE2D, were computed for each molecule using the top ranked 3D conformations generated by GOLD. These data were then used to predict the pIC_{50} values with SVR as described above (Table V). However, the results for the 1DSS, 2DSS and 1DSS + MOE2D descriptor libraries were found to be worse than those from the SVR models based on molecular conformations generated by CORINA (Table II). For MOE2Di3D, $r^2_{test} = 0.31$ was closer to the value 0.36 obtained for the original SVR model. This can be explained by the fact that configuration-dependent MOE i3D descriptors represent only 9% of the descriptor set.

DISCUSSION

As a first step towards developing global AChE models, we have evaluated a combination of Shape Signatures and MOE descriptors with PLS, and two SVM based methods, namely SVR and SVC. Our data using a benchmark AChE dataset (9) indicates that PLS and SVR with Shape Signatures and MOE descriptors are able to produce similar test set statistics as other widely accepted and more compute-intensive QSAR methods (e.g., CoMFA and CoMSIA). It is important to note that in this study we have used descriptors that are alignment independent. Sutherland *et al.* used alignment-dependent descriptors, which in turn were found to outperform alignment-independent descriptors, such as HQSAR and Cerius 2 (9). Therefore a more reliable comparison with our results would use the latter methods, in which case combining 1D Shape Signatures and MOE descriptors improves the test set r^2_{test} to 0.44 from 0.16–0.37 (Supplemental Table 1) in the study by Sutherland *et al.* (9).

The best overall regression model is the SVR paired with the 1DSS + MOE2D descriptor scheme. This model combines molecular shape expressed through the subset of 1DSS descriptors with various physicochemical characteristics captured in the MOE2D subset and was found to be superior to either shape-based alone or shape-free 2D MOE derived models. Based on our prior experience using classification studies (14,15,20), we expected the 2DSS/SVR model to outperform the 1DSS/SVR model. However, the 1DSS/SVR model with the Gaussian radial basis function kernel was preferable in modeling the binding affinity for the AChE

Table V. ε-SVR (support vector regression) Analysis of AChE Compounds from Sutherland *et al.* (9) with the 3D Descriptors Generated Using the Top Ranked GOLD Poses for Each Molecule. The Training Set Contains 73 Molecules and the Test Set Contains 37 Molecules

Molecular descriptors	s_{10CV}	r^2_{train}	s_{train}	r^2_{test}	s_{test}
1DSS	1.17	0.66	0.82	0.27	1.13
2DSS	1.06	0.66	0.76	0.17	1.20
MOE2Di3D	1.03	0.78	0.59	0.31	1.11
1DSS + MOE2D	1.05	0.64	0.83	0.33	1.08

Notations are the same as in Table II.

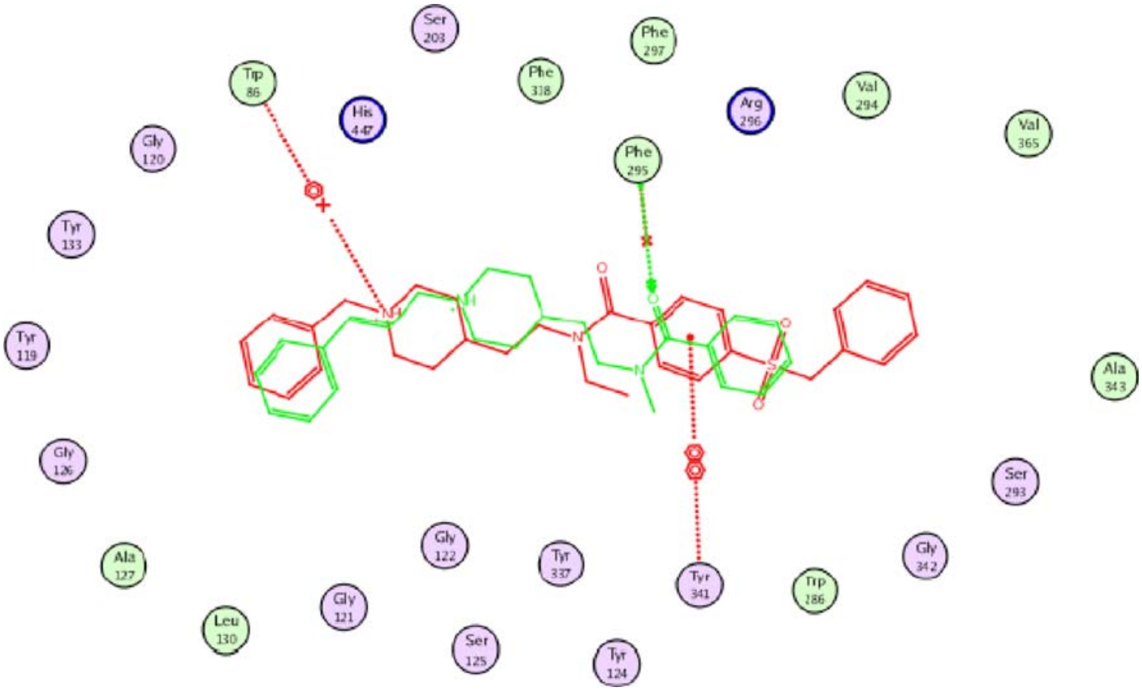

Fig. 2. Schematic representation of the interaction map and relative orientation inside the AChE binding pocket of Suth 2–22 (*red*) and Suth 2–32 (*green*) after docking with GOLD. Both compounds showed similar GOLD scores, 72.35 for Suth 2–22 and 67.78 for Suth 2–32, and were classified as strong inhibitors in the docking-based classification. Experimentally, only Suth 2–22 was found to be a strong binder with $IC_{50,exp}$=0.3 nM while Suth 2–32 was reported as a non-activator with $IC_{50,exp}$= 54 μM. The two molecules were assigned correctly by the 1DSS + MOE2D/SVM classifier. The binding site residues are colored by their nature, with hydrophobic residues in green, polar residues in purple and charged residues highlighted with bold contours. The figure was generated using the LigX application in MOE (Chemical Computing Group, Montreal, Canada).

receptor. It is also interesting to note that the MOE2D descriptors turned out to be the least sensitive to the choice of the regression algorithm.

In addition, we have used GOLD docking scores to classify compounds as inhibitors of AChE. This method was found to perform substantially worse than the SVM classification models coupled with the Shape Signatures and MOE descriptors. As was noted before, a weak correlation between the docking scores of the top GOLD poses and experimentally measured activities underlies the low predictive power of the docking based classification. This translated to a high rate of false predictions with 44% of the considered molecules being classified incorrectly. As an illustration, Fig. 2 displays the suggested interaction map and relative orientation inside the AChE binding pocket of the two molecules Suth 2–22 and Suth 2–32. The two structures produced similar GOLD scores, (72.35 for Suth 2–22, and 67.78 for Suth 2–32) and were found to be well positioned inside the binding pocket with a number of suggested stabilizing interactions to hold them in place. However, experimentally Suth 2–22 is a very strong inhibitor with $IC_{50,exp}$=0.3 nM, whereas Suth 2–32 is a weak binder ($IC_{50,exp}$=54 μM). Classification of the top

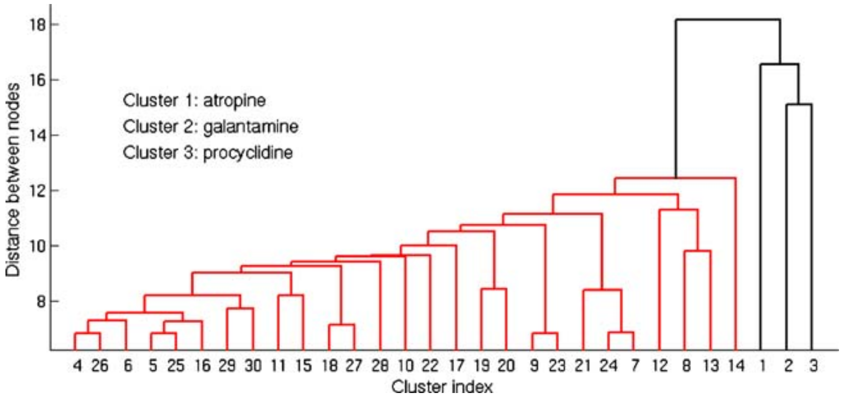

Fig. 3. The results of UPGMA (Unweighted Pair Group Method with Arithmetic Mean Algorithm) clustering analysis based on the Euclidean metric performed using the 1DSS + MOE2D descriptors for the Sutherland training dataset augmented by atropine, galantamine and procyclidine. UPGMA was performed using the Statistics Toolbox in MATLAB (Version 7.6, The MathWorks, Inc, Natick, MA).

docking modes of these molecules based on GOLD scores failed to reflect such a difference in activities. In contrast, both molecules are predicted correctly by the 1DSS + MOE2D/SVM classifier. This observation suggests that more detailed AChE-specific scoring functions are in fact needed to better represent ligand binding interactions of a given set of molecules with the AChE receptor using GOLD. A similar conclusion has been reported earlier by Guo *et al.* (39) for a different set of AChE inhibitors.

We have also addressed the question of which set of 3D molecular conformations is better suited for statistical modeling of the subject AChE inhibition data: the collection of the top ranked poses from GOLD (a single highest scored conformation for each molecule) or the ensemble of default low energy 3D conformations generated by CORINA (one conformation per structure). Intuitively, one might expect that the SVR models constructed from the top scoring GOLD conformations yield greater statistical quality than those obtained with structures produced by CORINA (Table II). However, we found the results with GOLD less satisfactory than those obtained with CORINA (Table V). A possible explanation of these findings may be that, in this particular case, the single top-ranked GOLD conformation may not be the best representation of the molecule. Perhaps building an averaged structure from the top 10% or so poses or utilizing a more rigorous scoring scheme for selecting a single conformation (39) would provide a more satisfactory description. In this context, CORINA assembles the molecular structure using a library of crystallographically determined structural templates, thus producing an averaged representation of the given molecule, which turned out to be better suited for the purposes of the reported statistical analysis.

Finally, we note the limitations of our approach, namely that not all known AChE molecules can be predicted with the PLS, SVC and SVR models based on the different descriptor schemes reported in this study. It is important that the users first ascertain whether their query molecules are inside the region of the chemical space (applicability domain) occupied by the molecules from the training set. For instance, we found that the known AChE inhibitors, viz., atropine, galantamine and procyclidine, cannot be reliably predicted with the models developed in this study from the published data from Sutherland *et al*. Fig. 3 illustrates the results of UPGMA (Unweighted Pair Group Method with Arithmetic Mean Algorithm) clustering analysis conducted in the space of 1DSS + MOE2D molecular descriptors on the Sutherland training dataset along with atropine, galantamine and procyclidine. This analysis clearly indicates that these molecules lie outside the training set and would explain their poor prediction using SVR.

In conclusion, the objective of the present study was to examine the quality of a novel set of alignment-independent molecular descriptors derived from molecular Shape Signatures (14,15,18–20). These descriptors are inherently three-dimensional and a relatively new addition to the other 2D/3D descriptor collections used in predictive QSAR modeling (40,41). These descriptors would also appear to be comparable to alignment-dependent descriptors used earlier by others with the same dataset. To date, we have shown that Shape Signatures can be used in drug discovery (19,21) and ADME/Tox models (14,15,20). We have for the first time extended the Shape Signatures methodology to regression and classification models for AChE inhibitors. Future work will focus on the development of scoring functions and machine learning-based methods that can capture the binding modes of structurally dissimilar ligands binding to AChE including CWAs. The Shape Signatures method is also currently under study as a novel approach for identification of simulants of CWAs by querying molecular databases. The models used in the current study could be used to score structurally similar molecules for therapeutic applications.

ACKNOWLEDGMENTS

We gratefully acknowledge all those involved in the application and development of Shape Signatures. Support for this work has been provided by the USEPA-funded Environmental Bioinformatics and Computational Toxicology Center (ebCTC), under STAR Grant number GAD R 832721-010, and by the Defense Threat Reduction Agency, under contract number HDTRA- BB07TAS020. This work was also funded by NIH -GM081394 from the National Institute of General Medical Sciences (to WJW). This work has not been reviewed by and does not represent the opinions of the funding agencies.

REFERENCES

1. Moretto A. Experimental and clinical toxicology of anticholinesterase agents. Toxicol Lett. 1998;102–103:509–13.
2. Castro A, Martinez A. Peripheral and dual binding site acetylcholinesterase inhibitors: implications in treatment of Alzheimer's disease. Mini Rev Med Chem. 2001;1:267–72.
3. Barril X, Orozco M, Luque FJ. Towards improved acetylcholinesterase inhibitors: a structural and computational approach. Mini Rev Med Chem. 2001;1:255–66.
4. Kaur J, Zhang MQ. Molecular modelling and QSAR of reversible acetylcholines-terase inhibitors. Curr Med Chem. 2000;7:273–94.
5. Cramer RD, Patterson DE, Bunce JD. Comparative Molecular Field Analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc. 1988;110:5959–67.
6. Tong W, Collantes ER, Chen Y, Welsh WJ. A comparative molecular field analysis study of N-benzylpiperidines as acetylcholinesterase inhibitors. J Med Chem. 1996;39:380–7.
7. Golbraikh A, Bernard P, Chretien JR. Validation of protein-based alignment in 3D quantitative structure-activity relationships with CoMFA models. Eur J Med Chem. 2000;35:123–36.
8. El Yazal J, Rao SN, Mehl A, Slikker W Jr. Prediction of organophosphorus acetylcholinesterase inhibition using three-dimensional quantitative structure-activity relationship (3D-QSAR) methods. Toxicol Sci. 2001;63:223–32.
9. Sutherland JJ, O'Brien LA, Weaver DF. A comparison of methods for modeling quantitative structure-activity relationships. J Med Chem. 2004;47:5541–54.
10. Fernandez M, Caballero J. Ensembles of Bayesian-regularized genetic neural networks for modeling of acetylcholinesterase inhibition by huprines. Chem Biol Drug Des. 2006;68:201–12.
11. Akula N, Lecanu L, Greeson J, Papadopoulos V. 3D QSAR studies of AChE inhibitors based on molecular docking scores and CoMFA. Bioorg Med Chem Lett. 2006;16:6277–80.
12. Jung M, Tak J, Lee Y, Jung Y. Quantitative structure-activity relationship (QSAR) of tacrine derivatives against acetylcholinesterase (AChE) activity using variable selections. Bioorg Med Chem Lett. 2007;17:1082–90.
13. Manchester J, Czermiński R. SAMFA: simplifying molecular descriptors for 3D-QSAR. J Chem Inf Model. 2008;48:1167–73.
14. Chekmarev DS, Kholodovych V, Balakin KV, Ivanenkov Y, Ekins S, Welsh WJ. Shape signatures: new descriptors for predicting cardiotoxicity in silico. Chem Res Toxicol. 2008;21:1304–14.

15. Kortagere S, Chekmarev D, Welsh WJ, Ekins S. New predictive models for blood-brain barrier permeability of drug-like molecules. Pharm Res. 2008;25:1836–45.
16. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. J Mol Biol. 1997;267:727–48.
17. Gasteiger J, Marsili M. Iterative partial equalization of orbital electronegativity–a rapid access to atomic charges. Tetrahedron 1980;36:3219–28.
18. Zauhar RJ, Moyna G, Tian L, Li Z, Welsh WJ. Shape signatures: a new approach to computer-aided ligand- and receptor-based drug design. J Med Chem. 2003;46:5674–90.
19. Nagarajan K, Zauhar R, Welsh WJ. Enrichment of ligands for the serotonin receptor using the Shape Signatures approach. J Chem Inf Model. 2005;45:49–57.
20. Kortagere S, Chekmarev D, Welsh WJ, Ekins S. Hybrid scoring and classification approaches to predict human pregnane X receptor activators. Pharm Res. 2009;26(4):1001-11.
21. Wang CY, Ai N, Arora S, Erenrich E, Nagarajan K, Zauhar R, et al. Identification of previously unrecognized antiestrogenic chemicals using a novel virtual screening approach. Chem Res Toxicol. 2006;19:1595–601.
22. Meek PJ, Liu Z, Tian L, Wang CY, Welsh WJ, Zauhar RJ. Shape Signatures: speeding up computer aided drug discovery. Drug Discov Today. 2006;11:895–904.
23. Kortagere S, Welsh WJ. Development and application of hybrid structure based method for efficient screening of ligands binding to G-protein coupled receptors. J Comput Aided Mol Des. 2006;20:789–802.
24. Whitley DC, Ford MG, Livingstone DJ. Unsupervised forward selection: a method for eliminating redundant variables. J Chem Inf Comput Sci. 2000;40:1160–8.
25. Geladi P, Kowalski B. Partial least-squares:a tutorial. Anal Chim Acta. 1986;185:1–17.
26. Cortes C, Vapnik V. Support vector networks. Machine Learn. 1995;20:273–93.
27. Vapnik V. Statistical learning theory. New York: Wiley; 1998.
28. Kecman V. Learning and soft computing: support vector machines, neural networks and Fuzzy logic models. Cambridge: MIT; 2001.
29. Ivanciuc O. Application of support vector machines in chemistry. Rev Comp Chem. 2007;23:291–400.
30. Chen YZ, editor. Current QSAR techniques for toxicology. Hoboken: Wiley; 2007.
31. Xue Y, Yap CW, Sun LZ, Cao ZW, Wang JF, Chen YZ. Prediction of P-glycoprotein substrates by a support vector machine approach. J Chem Inf Comput Sci. 2004;44:1497–505.
32. Leong MK. A novel approach using pharmacophore ensemble/support vector machine (PhE/SVM) for prediction of hERG liability. Chem Res Toxicol. 2007;20:217–26.
33. Ung CY, Li H, Yap CW, Chen YZ. In silico prediction of pregnane X receptor activators by machine learning approaches. Mol Pharmacol. 2007;71:158–68.
34. Song M, Breneman C, Bi J, Sukumar N, Bennett K, Cramer S, et al. Prediction of protein retention times in anion-exchange chromatography systems using support vector regression. J Chem Inf Compu Sci. 2002;42:1347–57.
35. Yap CW, Li ZR, Chen YZ. Quantitative structure-pharmacokinetic relationships for drug clearance by using statistical learning methods. J Mol Graph Model. 2006;24:383–95.
36. Chang CC, Lin CJ. LIBSVM: a library for support vector machines, 2001. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.
37. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. 1975;405:442–51.
38. Kryger G, Harel M, Giles K, Toker L, Velan B, Lazar A, *et al*. Structures of recombinant native and E202Q mutant human acetylcholinesterase complexed with the snake-venom toxin fasciculin-II. Acta Crystallogr Sect D. 2000;56:1385–94.
39. Guo J, Hurley MH, Wright JB, Lushington GH. A docking score function for estimating ligand-protein interactions: application to acetylcholinesterase inhibition. J Med Chem. 2004;47:5492–500.
40. Ekins S, Embrechts MJ, Breneman CM, Jim K, Wery J-P. Novel applications of Kernel-partial least squares to modeling a comprehensive array of properties for drug discovery. In: Ekins S, editor. Computational toxicology: risk assessment for pharmaceutical and environmental chemicals. Hoboken: Wiley-Interscience; 2007. p. 403–32.
41. Todeschini R, Consonni V. Handbook of molecular descriptors. Weinheim: Wiley-VCH; 2000.