SAMFA: Simplifying Molecular Description for 3D-QSAR

John Manchester and Ryszard Czerminski*

AstraZeneca Pharmaceuticals R&D Boston, 35 Gatehouse Drive, Waltham, Massachusetts 02451

Received January 10, 2008

In this paper we consider the following question: How much can we simplify molecular description without sacrificing too much quality of 3D-QSAR models. We compare the performance of the newly developed Simple Atom Mapping Following Alignment (SAMFA) descriptors with CoMFA using nine different data sets from the literature, by using three regression approaches (PLS, SVM, RandomForest), as implemented in R, and Monte Carlo cross-validation (MCCV) numerical experiments. The results indicate that SAMFA descriptors, despite their simplicity, perform surprisingly well when compared to the much more refined CoMFA descriptors. Moreover, their simplicity makes them readily interpretable and applicable to the difficult problem of inverse QSAR.

"Make everything as simple as possible, but not simpler." - Albert Einstein

1. INTRODUCTION

Quantitative structure−activity relationships (QSAR) represent an attractive approach for predicting compound activities when compared to more elaborate computational approaches (e.g., linear interaction energy approach[6]), with the advantages being speed and low cost.

Since the publication of CoMFA (Comparative Molecular Field Analysis) in 1988,[1] literally hundreds of studies have appeared applying the technique with remarkable success. The originality of CoMFA lies in representing molecules by their steric and electrostatic fields sampled on a regular grid to produce a large number of descriptors that are conveniently handled using partial least-squares (PLS).[8] Despite the introduction of many subsequent 3D-QSAR techniques, aimed explicitly at addressing shortcomings of CoMFA (already delineated in the original paper) and seeking more sophisticated treatment of the molecular field or how it is sampled, the method remains one of the most robust and predictive.[9]

Molecular fields however, although conceptually attractive, in practice can be difficult to interpret for medicinal chemists, who are used to thinking in literal terms of what functional groups are acceptable at various positions on a given scaffold. In addition, as has been extensively discussed in the original CoMFA[1] and CoMSIA[10] papers, molecular field descriptors are susceptible to small changes in compound alignment as well as artifacts arising from sampling on a regular 3-D grid. Additionally, when fields are used, information from individual atoms and molecular features is convoluted. It is, therefore, not possible to reconstruct specific molecules from fields (inverse QSAR problem[11]).

Here, we explore what happens if a significantly *less* sophisticated description is used for the molecules. To this end we employ an approach inspired by Free and Wilson methodology,[12] and we use as QSAR descriptors atom types of atoms located in specific 3D locations, from compounds aligned in the same way as for the CoMFA method.

It turns out that the approach performs surprisingly well in building predictive models. We thus somewhat facetiously named the approach SAMFA (**S**imple **A**tom **M**apping **F**olowing **A**lignment). To rigorously compare the performance of QSAR models built using SAMFA and CoMFA descriptors we applied both approaches to nine literature data sets (the steroid benchmark,[13] ACE, AChe, BZR, COX2, DHFR, GPB, THERM, and THR (the last eight compiled by Sutherland[9])) and found that SAMFA performance is comparable or even somewhat better than CoMFA.

We attribute this interesting finding to a combination of modeling approximations and experimental errors, which results in the SAMFA description, despite its simplicity, being good enough or "satisficing".[14,15] The word satisfice was coined by Herbert A. Simon[16] as a portmanteau of "satisfy" and "suffice".

2. METHODOLOGY

2.1. MCCV: Monte Carlo Cross-Validation. Monte Carlo cross-validation (MCCV)[3–5] (also known as random subsampling or multiple hold-out) was used to assess the predictive ability of the regression models. In MCCV the whole data set is partitioned randomly into two parts. One part (train sample) is used to build the model, the remaining part (test sample) is used for prediction, and the whole process is repeated until the convergence of $median(q^2)$ is achieved. Here we use 10% of the whole data set as a test sample and denote this accordingly as $MCCV_{10}$. MCCV is more time-consuming but yields a more robust and detailed assessment of model quality than the often used leave-one-out (LOO) cross-validation method or "single hold-out set validation", which is sensitive to idiosyncrasies of the train/test partition−in particular for small data sets (e.g., steroids). As the main measure of quality of the models we use $q^2 = 1 - \mathrm{MSE}/\mathrm{Var}(Y)$, with

* Corresponding author phone: (781)839-4304; fax: (781)839-4220; e-mail: ryszard.czerminski@astrazeneca.com.

10.1021/ci800009u CCC: $40.75
Published on Web 05/27/2008

$$\mathrm{MSE} = \sum_{i=1}^{N} (Y_i^{\mathrm{obs}} - Y_i^{\mathrm{pred}})^2/N$$

and

$$\mathrm{Var}(Y) = \sum_{i=1}^{N} (Y_i^{\mathrm{obs}} - \mathrm{mean}(Y^{\mathrm{obs}}))^2/N$$

where the response variable $Y = \mathrm{pIC50} = -\log_{10}(\mathrm{IC50[M]})$ (or $\mathrm{p}K_i$ for some data sets, as indicated). Data set partitions which resulted in test samples with $\mathrm{Var}(Y) = 0$ were discarded in order to avoid division by zero.

From MCCV results we can get an estimate of q^2 probability density (p) and cumulative distribution function

$$\Pr(q^2 < x) = \int_{-\infty}^{x} p(t)\mathrm{d}t$$

This is illustrated in Figure 1. Density distribution can be conveniently characterized by *median*(q^2) (horizontal heavy line segment), Pr ($q^2 > 0.5$) (vertical heavy line segment), and *integral* q^2

$$q^2.\mathrm{int} = \int_0^1 [1 - \Pr(q^2 < x)]\mathrm{d}x \qquad (1)$$

which is simply an area *above* $\Pr(q^2<x)$ for $0 < x < 1$ as illustrated by the shaded area in Figure 1b. *Integral* q^2 (q^2.int) represents an overall measure of model quality for a given descriptor/method combination: the value of one corresponds to a situation where all models from MCCV generate $q^2 = 1$ results, and the value of zero corresponds to the other extreme where all MCCV models yield $q^2 \leq 0$; for all realistic models we should have $0 < q^2.\mathrm{int} < 1$.

All regression results were obtained with PLS (pls.pcr package), RandomForest (randomForest package), and SVM (e1071 package) methods as implemented in the R environment[2] (R version 2.5.1 (2007-06-27) on Linux). For PLS the default 10-fold cross-validation was used to define the number of latent variables to use, the number of trees in RandomForest was set to the default value 500, and the RBF kernel was used for SVM with the default parameters ($\gamma = 1/$*ndescriptors* and $C = 1$). All descriptor sets were autoscaled (columns mean centered and divided by standard deviation) prior to their use in regression methods. We did not perform any descriptor selection since all three methods we are using here (PLS, SVM, and RandomForest) employ some form of regularization[17] and are designed to work well for ill-defined problems (more descriptors then molecules), although the SVM approach seems to be somewhat sensitive to the presence of "noise" variables.[18] For Monte Carlo simulation to be useful it is important to ascertain that results are stable. Convergence of *median*(q^2) as a function of the sample size is illustrated in Figure 2, in which the mean of *median*(q^2) (together with standard deviation) from 1024 samples out of 2^{13} independent MCCV_{10} runs, with the size of the samples varying from 2^5 to 2^{12} is plotted. Numerical experiments indicate that 2^{12} repetitions are enough to generate convergent *median*(q^2) values for all data sets.

2.2. SAMFA: Simple Atom Mapping Following Alignment. In general, an alignment rule must first be applied to the set of compounds as in other grid-based approaches. Instead of a *regular* grid, however, an *irregular* template is constructed based on the aligned compounds. In Coats' data set,[13] the ten additional steroids originally identified by Cramer as an external test set are aligned in a separate

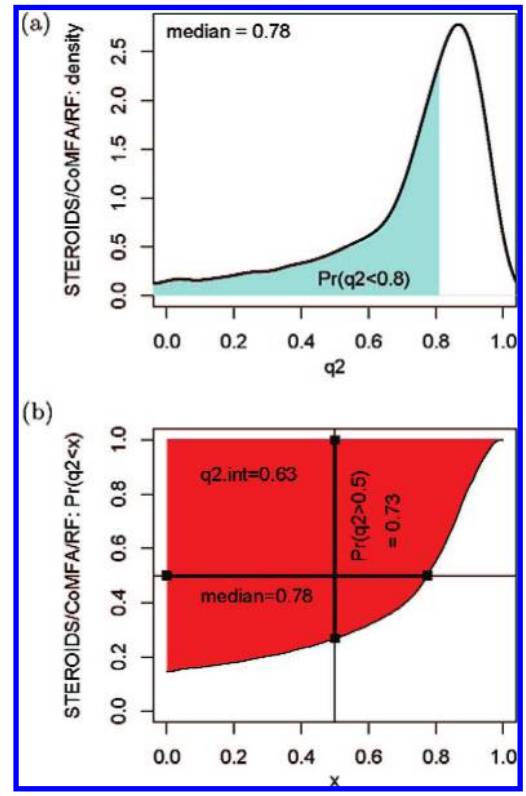

Figure 1. STEROIDS/CoMFA/RF: (a) q^2 probability density (p) and (b) cumulative probability (Pr): shaded area represents q^2int as defined in eq 1.

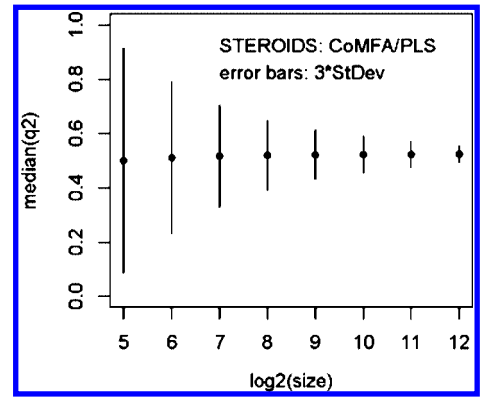

Figure 2. Convergence of *median*(q^2) values as a function of sample size for the most slowly converging MCCV_{10} simulation. Vertical lines indicate $\pm 3\sigma$.

reference frame. We realigned the entire data set on deoxycortisol (steroid 11) with the slight variation of aligning each molecule on the maximum common substructure with deoxycortisol instead of using carbon atoms 3, 5, 6, 13, 14, and 17 as in the original CoMFA paper. In-house program jb_mcss,[19] was used for this task. This variation does not significantly change the alignment of the original 21 steroids, and we used it simply for convenience. For Sutherland data sets[9] we used the alignments provided in the Supporting Information.

The initial template is taken as the coordinates of all atoms in the largest compound in the set; that is, all atoms are represented as dummy atoms, and bonding information is discarded. Each remaining compound is then considered in turn, and if any atom in the compound lies beyond a predetermined cutoff distance from all dummy atoms in the template, then a new dummy atom is added to the template

with the coordinates of that atom. Finally the dummy atoms in the template are assigned identifiers. Following this procedure, the template is guaranteed to possess one unique *irregular* grid point for each atom among the molecules in the aligned data set, thus eliminating noise arising from *rectangular* grid artifacts. In our hands satisfactory results were obtained using a cutoff of 1.2 Å, and we found results to be not very sensitive to the cutoff value, e.g. 1.4 Å cutoff results in only a slightly worse performance.

Once the template is complete, a fingerprint is constructed by creating variable names for each combination of dummy atom identifier and allowable atom type. The allowable atom types used in the present work are C, N, O, S, P, F, Cl, Br, I, and H. In addition, we found that allowing generalization of atom types to pharmacophoric elements, such as "hydrogen bond donor", improved the performance of the method. Therefore, we additionally allowed atoms to assume the following identities: X (halogen); aroC (aromatic C), aliC (aliphatic C); HBA (hydrogen bond acceptor); HBD (hydrogen bond donor); and EWG (electron-withdrawing group). The use of an Electron-Donating Group (EDG) descriptor was not explored. Separate detection is performed for nitriles, which are mapped as EWGs at the carbon position. Mapping is done for each molecule in the data set by finding the closest dummy atom in the template and flipping the appropriate bits in the fingerprint. For example, if a fluorine atom corresponding to template dummy atom Du123 is encountered, the bits Du123_F, Du123_X, and Du123_EWG are set to true. Finally, only polar hydrogen atoms are recorded (i.e., those bonded to HBD atoms) in order to retain possible information about directionality of hydrogen bonds to the receptor, while avoiding noise associated with rotation of methyl groups, etc. For computational efficiency, bits with null variance for the entire data set are excluded before outputting the final fingerprints. The SAMFA approach was implemented using the OEChem library developed by OpenEye.[20]

2.3. CoMFA: Comparative Molecular Field Analysis. CoMFA uses as a descriptors steric and electrostatic field values sampled on a rectangular grid surrounding prealigned molecules and applies PLS as a regression method.[1]

CoMFA implementation in SYBYL (version 7.3 on Linux platform) (Tripos Inc.; St. Louis, MO) provides a number of adjustable parameters. We used the following: default grid spacing (2 Å) and centering, spherical smoothing, Drop Electrostatic: "Never". Interestingly, by default in SYBYL electrostatic terms are dropped: "Within Steric Cutoff for Each Row", which introduces some deterioration of LOO-q^2 values. As demonstrated in ref 21 with extensive optimization of CoMFA settings it is possible to obtain better models (as measured by LOO-q^2 and r^2_{pred}) for individual data sets. However, optimal settings are not consistent between different data sets, and there are many instances where optimal settings for one data set result in models performing worse than the default for a different data set—a hallmark of overfitting. We therefore restricted our modifications to the default settings as described above.

3. DATA SETS

The steroid benchmark data set has been used extensively in 3-D QSAR validation studies and, therefore, scrutinized by numerous investigators. In reviewing this literature, Coats[13] compiled the noted idiosyncrasies of the data set and in some cases corrected errors. He has made the "corrected" data set available, which we have used in the present work. Sutherland et al.[9] have similarly reviewed the literature, compiled a set of eight other data sets, and made these available. These are as follows: ACE: pIC50 for angiotensin converting enzyme−inhibitors; ACHE: pIC50 for acetylcholinesterase inhibitors; BZR: pIC50 for benzodiazepine receptor ligands; COX2: pIC50 for cyclooxygenase-2 inhibitors; DHFR: pIC50 for dihydrofolate inhibitors; GPB: pK_i for glycogen phosphorylase inhibitors; THER: pK_i for thermolysin inhibitors; and THR: pK_i for thrombin inhibitors. Basic data set statistics are given in Table 1. More details about the data sets can be found in Sutherlands' compilation[9] and in original papers cited therein.

Table 1. Basic Data Set Statistics: Number of Molecules and Min/Mean/Median/Max/Var for the Response Variable (pIC50 or pK_i) together with the Number of Independent SAMFA and CoMFA Descriptors[c]

	N	Min	Mean	Median	Max	Var	ND[a]	ND[b]
STEROIDS	31	5.00	6.38	6.28	7.88	1.17	117	1600/1555
ACE	114	2.14	6.32	6.39	9.94	5.20	888	4004/3999
ACHE	111	4.27	6.78	6.84	9.52	1.54	438	4368/4336
BZR	163	5.00	7.50	7.80	8.92	1.21	394	2860/2837
COX2	322	4.00	6.43	6.60	9.00	2.03	530	2880/2843
DHFR	397	3.30	5.90	6.06	9.81	2.21	854	2904/2885
GPB	66	1.30	2.77	2.45	6.80	1.26	271	2376/2348
THERM	76	0.52	5.02	5.36	10.17	4.11	563	4680/4674
THR	88	4.36	6.65	6.69	8.48	1.02	478	3432/3381

[a] SAMFA descriptors. [b] CoMFA descriptors. [c] For CoMFA the number of descriptors before and after removing columns with zero variance is shown. There were no null variance columns in the SAMFA descriptors. The STEROIDS data set is from Coats;[13] all others are from Sutherland.[9]

4. RESULTS

In this study we compare the performance of simple 3D descriptors (SAMFA) with CoMFA using nine literature data sets, three regression methods, and MCCV (full q^2 probability distribution) and LOO cross-validated q^2 values. In most cases CoMFA and SAMFA performance is equivalent for all practical purposes, and in the GPB case SAMFA performance seems to be somewhat better, as illustrated in Figure 3. In generating these plots, for each data set/descriptor combination the results of the best performing regression method were used to minimize regression bias. However, differences between different regression methods do not change the conclusions in any qualitative fashion. In most cases the RandomForest (RAF) regression method delivers the best performance, and in cases when SVM or PLS generate better models, the RAF models have very similar quality (see Tables 2 and 3). Note (Figure 4) that *median*(q^2) from MCCV_{10} correlates strongly with the LOO-CV q^2 values, which is intuitively pleasing. An advantage of using MCCV over single points estimate of model quality (like LOO-q^2) is that it generates an estimate of probability distribution of q^2 values and therefore allows for making informed statements about getting q^2 values in a certain range e.g. $\Pr(q^2 < 0) = 0.2$ (Figure 1).

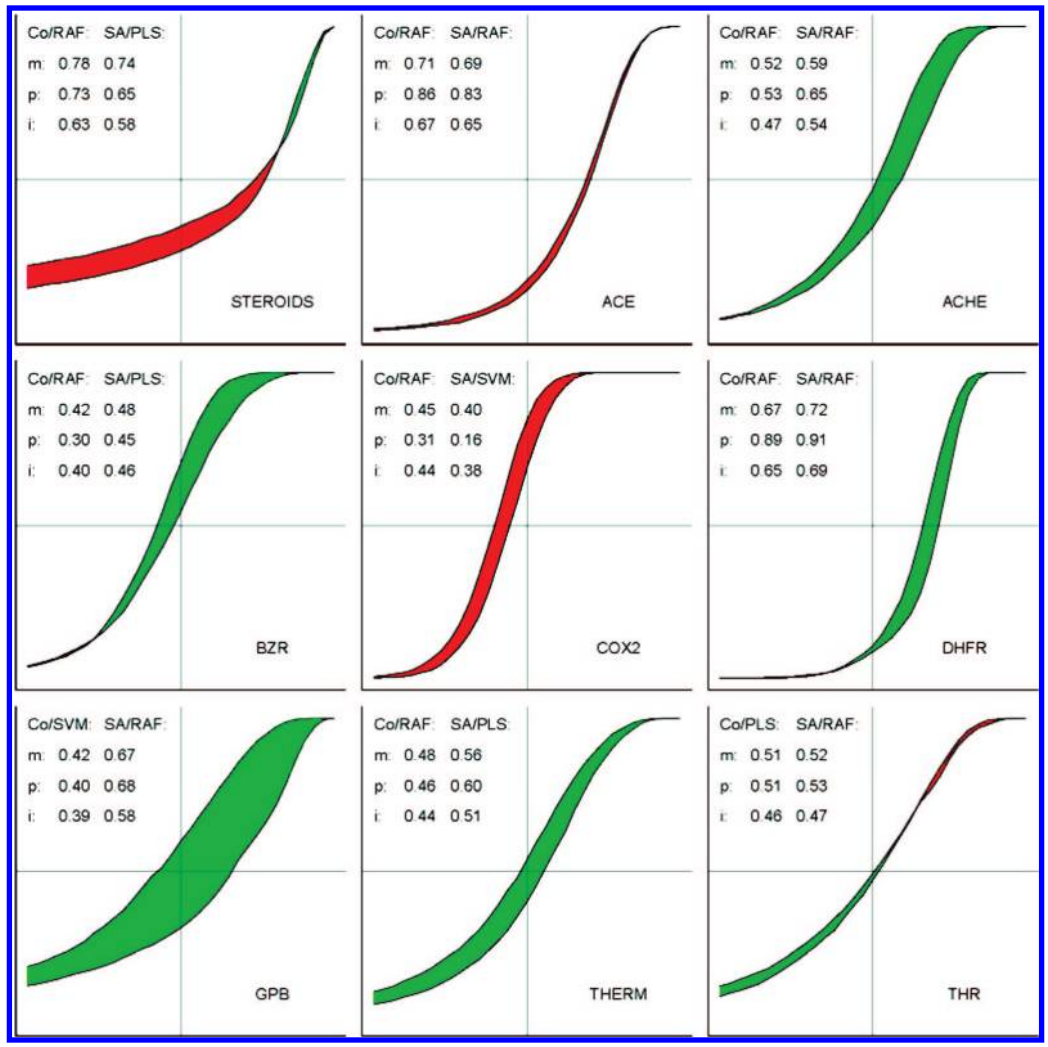

Figure 3. Comparison between SAMFA and CoMFA performance for all nine data sets using cumulative probability: $\Pr(q^2 < x)$; see Figure 1. For each data set $MCCV_{10}$ results of best performing regression methods were used: *m, median*(q^2); *p*, $\Pr(q^2 > 0.5)$; *i*, q^2.int. Red shaded area reflects better performance for CoMFA on the data set (cumulative probability curve of CoMFA below SAMFA); green (or lighter gray in black and white) indicates better performance for SAMFA.

5. SAMFA: MODEL INTERPRETATION

Interpretation is straightforward and involves simply assigning the appropriate atom types to particular dummy atoms in the template found to correlate (or anticorrelate) with activity as determined by the model. A very effective interpretation can be achieved by viewing the resulting "consensus molecules" together with representative compounds from the data set in a standard molecular graphics visualizer such as PyMOL[22] or VIDA.[20] We use the STEROIDS data set[13] for illustration.

For example, Figure 5a shows the SAMFA descriptors correlated with enhanced corticosteroid-binding globulin (CBG) binding superimposed on cortisol, which exhibits high CBG affinity. These features are consistent with the SAR of the steroid benchmark. Seventeen β-OH (SAMFA H-bond donor), an aromatic A-ring, and an sp^2-hybridized C6 are correlated with decreased CBG affinity in the steroid data set.[23] Interestingly, a number of additional features highlighted by SAMFA are also consistent with interactions observed in the recently determined X-ray crystal structure of cortisone-bound rat CGB.[24] The 11 β-OH, captured by SAMFA as an H-bond donor off the C11 position on steroid ring C, interacts with the backbone of residue 256 and the side chain of K260 (N260 in human). Comparison with the SAMFA descriptors anticorrelated with CGB affinity (Figure 5b) reveals that a H-bond acceptor at this position decreases CBG affinity (as in 4-pregnene-3,11,20-trione). Similarly, the presence of part or all of the β-hydroxy ketone C17 substituent on ring D is associated with enhanced affinity, consistent with the observed interactions of groups in this region with a conserved Glu (Q224) in the crystal structure. Finally, the C3 carbonyl group is present in the good binders, while hydroxyl at this position abolishes affinity. The closest residue in rat is Pro14, which makes no polar interactions with cortisol in the crystal structure. In human this residue is a serine, which could donate an H-bond to the C3-substituent, consistent with the observed preference for a H-bond acceptor over a donor at C3.

In addition to simplicity, the big advantage of this way of interpretation is that it shows literally the functional groups that are correlated with improving activity. Medicinal chemists can in this way immediately understand the correlation structure within models and, by superimposing the significant features (SAMFA descriptors) on a compound of interest, begin to design new analogs of that compound to improve potency. This is an important

Table 2. Summary Data for q^2 Values;[a]

project	regression	descriptor	q^2.LOO	q^2.median	Pr(q^2 > 0.5)	q^2.int	nL.mean
STEROIDS	RF	CoMFA	0.77	0.78	0.73	0.63	
STEROIDS	RF	SAMFA	0.69	0.73	0.62	0.56	
STEROIDS	SVM	CoMFA	0.61	0.64	0.60	0.53	
STEROIDS	SVM	SAMFA	0.60	0.60	0.61	0.50	
STEROIDS	PLS	CoMFA	0.57	0.52	0.51	0.45	2.8
STEROIDS	PLS	SAMFA	0.69	0.74	0.65	0.58	4.3
ACE	RF	CoMFA	0.70	0.71	0.86	0.67	
ACE	RF	SAMFA	0.69	0.69	0.83	0.65	
ACE	SVM	CoMFA	0.55	0.57	0.66	0.53	
ACE	SVM	SAMFA	0.52	0.53	0.55	0.48	
ACE	PLS	CoMFA	0.66	0.62	0.72	0.58	5.2
ACE	PLS	SAMFA	0.65	0.63	0.75	0.59	3.9
ACHE	RF	CoMFA	0.51	0.52	0.53	0.47	
ACHE	RF	SAMFA	0.58	0.59	0.65	0.54	
ACHE	SVM	CoMFA	0.37	0.37	0.24	0.35	
ACHE	SVM	SAMFA	0.29	0.29	0.07	0.27	
ACHE	PLS	CoMFA	0.48	0.35	0.28	0.34	5.2
ACHE	PLS	SAMFA	0.54	0.51	0.51	0.47	3.4
BZR	RF	CoMFA	0.42	0.42	0.30	0.40	
BZR	RF	SAMFA	0.43	0.44	0.40	0.43	
BZR	SVM	CoMFA	0.36	0.37	0.14	0.36	
BZR	SVM	SAMFA	0.38	0.36	0.20	0.36	
BZR	PLS	CoMFA	0.42	0.36	0.22	0.33	5.5
BZR	PLS	SAMFA	0.49	0.48	0.45	0.46	4.0
COX2	RF	CoMFA	0.44	0.45	0.31	0.44	
COX2	RF	SAMFA	0.38	0.38	0.18	0.37	
COX2	SVM	CoMFA	0.44	0.44	0.33	0.43	
COX2	SVM	SAMFA	0.39	0.40	0.16	0.38	
COX2	PLS	CoMFA	0.40	0.39	0.18	0.37	7.1
COX2	PLS	SAMFA	0.40	0.40	0.22	0.38	3.0
DHFR	RF	CoMFA	0.66	0.67	0.89	0.65	
DHFR	RF	SAMFA	0.70	0.72	0.91	0.69	
DHFR	SVM	CoMFA	0.63	0.62	0.87	0.62	
DHFR	SVM	SAMFA	0.57	0.56	0.76	0.56	
DHFR	PLS	CoMFA	0.66	0.66	0.91	0.65	8.9
DHFR	PLS	SAMFA	0.68	0.69	0.90	0.67	5.9
GPB	RF	CoMFA	0.41	0.40	0.37	0.37	
GPB	RF	SAMFA	0.66	0.67	0.68	0.58	
GPB	SVM	CoMFA	0.40	0.42	0.40	0.39	
GPB	SVM	SAMFA	0.53	0.52	0.54	0.48	
GPB	PLS	CoMFA	0.47	0.34	0.36	0.34	6.1
GPB	PLS	SAMFA	0.61	0.58	0.60	0.49	3.5
THERM	RF	CoMFA	0.48	0.48	0.46	0.44	
THERM	RF	SAMFA	0.52	0.53	0.54	0.47	
THERM	SVM	CoMFA	0.32	0.34	0.25	0.32	
THERM	SVM	SAMFA	0.18	0.21	0.02	0.20	
THERM	PLS	CoMFA	0.44	0.35	0.29	0.33	3.3
THERM	PLS	SAMFA	0.60	0.56	0.60	0.51	7.3
THR	RF	CoMFA	0.46	0.44	0.39	0.41	
THR	RF	SAMFA	0.53	0.52	0.53	0.47	
THR	SVM	CoMFA	0.46	0.46	0.44	0.42	
THR	SVM	SAMFA	0.39	0.39	0.25	0.35	
THR	PLS	CoMFA	0.55	0.51	0.51	0.46	5.9
THR	PLS	SAMFA	0.56	0.48	0.47	0.44	5.7

[a] The mean value for the number of components in PLS regression is shown in the nL.mean column.

difference from field-based methods. While those methods can be used to produce appealing visualizations of virtual binding pockets, they suggest no information about specific molecular features and instead require the enumeration and virtual screening of libraries of compounds designed using relatively vague requirements of polarity and steric bulk around particular regions of the molecular scaffold.

6. DISCUSSION

In this paper SAMFA was compared to CoMFA using nine data sets. Common molecular alignments were used in calculating descriptors from each approach, and in order to use identical statistical treatment, the CoMFA descriptors were extracted from Sybyl. To minimize the effects of model bias, three different regression methods were applied, and to avoid potential idiosyncrasies arising from particular divisions of the data into training and test sets, a MCCV procedure was used to generate the probability density of q^2 values arising from each of the resulting combinations, which provides for a complete picture of the model quality that is likely to be achieved using a particular regression method on a given descriptor set. Comparing the best-performing regression methods for

Table 3. Statistics for q^2.int over All Data Sets Using Table 2 Values

descriptor/regr	Min	first Qu	Median	Mean	third Qu	Max
CoMFA/RAF	0.37	0.41	0.44	0.50	0.63	0.67
CoMFA/SVM	0.32	0.36	0.42	0.44	0.53	0.62
CoMFA/PLS	0.33	0.34	0.37	0.43	0.46	0.65
CoMFA/all	0.32	0.36	0.43	0.45	0.53	0.67
SAMFA/RAF	0.37	0.47	0.54	0.53	0.58	0.69
SAMFA/SVM	0.20	0.35	0.38	0.40	0.48	0.56
SAMFA/PLS	0.38	0.46	0.49	0.51	0.58	0.67
SAMFA/all	0.20	0.41	0.48	0.48	0.56	0.69
CoMFA/RAF	0.37	0.41	0.44	0.50	0.63	0.67
SAMFA/RAF	0.37	0.47	0.54	0.53	0.58	0.69
all/RAF	0.20	0.43	0.47	0.50	0.58	0.69
CoMFA/SVM	0.32	0.36	0.42	0.44	0.53	0.62
SAMFA/SVM	0.20	0.35	0.38	0.40	0.48	0.56
all/SVM	0.17	0.32	0.39	0.40	0.49	0.62
CoMFA/PLS	0.33	0.34	0.37	0.43	0.46	0.65
SAMFA/PLS	0.38	0.46	0.49	0.51	0.58	0.67
all/PLS	0.15	0.36	0.46	0.45	0.53	0.67

each data set for SAMFA vs CoMFA, we found that, on nine data sets studied, SAMFA is as good or slightly better than CoMFA.

We find it remarkable that the simple description used by SAMFA provides for QSAR models with performance comparable to the much more sophisticated CoMFA approach. We take this interesting result as a reminder of the relative crudeness of models compared to the complexity of the biological systems they are intended to approximate. Taken in the context of the simplifications made in 3D-QSAR—representing flexible ligands with single conformers aligned in an often arbitrary mode with little or no knowledge of the receptor, "conformer focusing",[25] exclusion of solvent effects, and treatment of the entire system, including derivation of complex molecular fields, using empirical force fields, to name a few—it is perhaps not at all surprising that the resulting model quality is somewhat insensitive to the form of descriptors used. In fact, because the molecular fields are derived from the molecular mechanics atom types in CoMFA, it is possible that the benefits of expanding the atoms into full molecular fields is off-set by the dilution of signal, or "the curse of dimensionality",[26] arising from introducing a few thousand variables from sampling that field on a Cartesian grid, not to mention grid artifacts which have been well documented by others.[1]

It is important to keep in mind that SAMFA shares some limitations with CoMFA, in particular it depends on alignment on the conformations in a common frame and might perform poorly when presented with molecules possessing structural elements, which are not "covered" in a training set. We are aware of the fact that CoMSIA, due to Gaussian smoothing of the fields, mitigates grid artifacts to some extent, but unfortunately we were not able to extract CoMSIA fields form SYBYL tables, and we do not expect CoMSIA results to be qualitatively different from CoMFA on these data sets.[9]

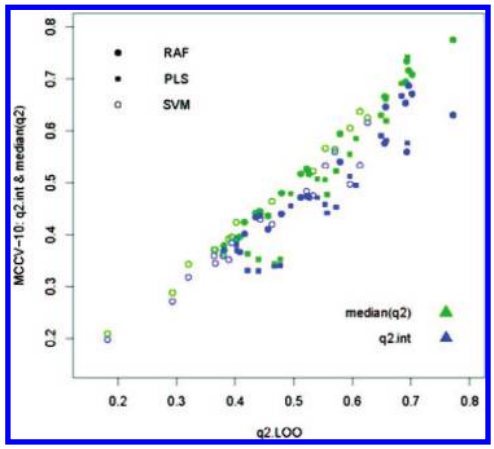

Figure 4. Relationships between q^2 values from LOO and $MCCV_{10}$ experiments.

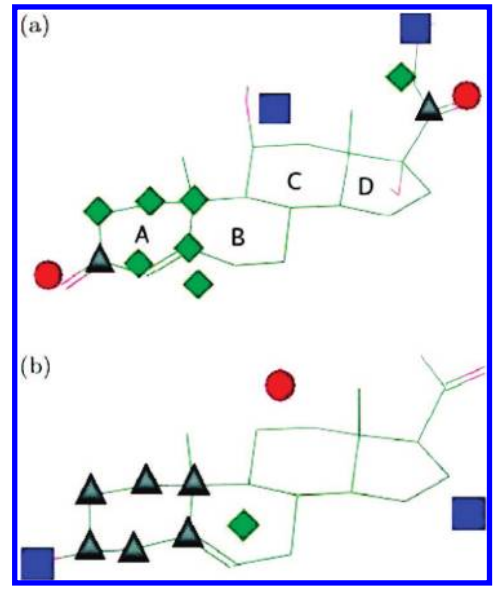

Figure 5. Interpretation of SAMFA descriptors is straightforward and amounts to extracting dummy atom/atom type combinations and displaying them in a visualization program. This has been done for features (a) correlated with enhanced CBG binding in the steroid data set; the high-affinity cortisol is shown as a reference compound. (b) Anticorrelated with CBG binding are shown with pregnenolone as the reference. Red circles are H-bond acceptors; blue squares are H-bond donors; green diamonds are sp3-carbons; and black triangles are sp2-carbons. Note that symbols indicate features overall correlated with activity and do not necessarily correspond to the atom types in any particular molecule.

Tirado-Rives et al.[25] carefully document errors in "conformer focusing", one of the simplest components in binding free energy prediction. They demonstrate that with current state-of-the-art methods, we can expect errors on the order of 5 kcal/mol (or larger) for drug-sized molecules. Considering all the various terms that are neglected in the much simpler 3D-QSAR treatment, it is remarkable that any success can be achieved at all.

It strikes us that the success of 3D-QSAR methods is not due to the sophistication of the underlying theoretical methods but rather to the fact that the various perturbations to free energy that result from a ligand binding a receptor can be accounted for in the parametrization of the regression models, particularly within congeneric series where cancelation of the largest errors is most likely. It of course helps that the descriptors have something to do with the structures of the compounds of interest, and using the molecular fields of compounds is intuitive and satisfying. But within a QSAR formalism, molecular fields seem to offer little, if any, advantage over what amounts to the force-field atom types used to derive those fields. Perhaps this observation is a reminder that before applying the most sophisticated methods at hand to a particular problem, the simplest approach should first be sought.

7. CONCLUSIONS

Based on the comparison between SAMFA and much more sophisticated CoMFA descriptors, we conclude that the SAMFA level of description is satisfying with respect to the predictive ability of QSAR modeling. We attribute this phenomenon to the number of simplifications and approximations involved in QSAR modeling resulting in a situation in which introduction of another simplification due to SAMFA does not have a detrimental effect on the predictive ability of QSAR models and in some cases might even allow for generating models with better predictive ability. As the most important advantage of a SAMFA-type model we see the ease of model interpretation.

ACKNOWLEDGMENT

We thank Dr. William Curtis (Tripos) for providing us with SPL script to extract CoMFA descriptors from SYBYL tables and Bob McLaughlin for discussion and occasional blues. We would like also to thank Reviewer #2 for pointing to the fact that Sutherland[9] reports q^2_{LOO} performance of the model generated using HQSAR descriptors to be comparable to CoMFA. This is interesting since HQSAR descriptors do not require overlay of the molecules. In this work however we wanted to focus on one-to-one comparison of SAMFA with current de facto standard in 3D QSAR field i.e. CoMFA. Another important aspect of SAMFA model is the ease of interpretation of the model; HQSAR or other similar in nature descriptors do not allow for such straightforward interpretation.

Supporting Information Available: We include source code for SAMFA approach to facilitate, for interested readers, experimentation with any other potential simplifications of molecular description in 3D-QSAR. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.
(2) R Development Core Team, R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2008; ISBN 3-900051-07-0. http://www.r-project.org (accessed Feb 11, 2008).
(3) Picard, R. R.; Cook, R. D. Cross-Validation of Regression Models. *J. Am. Stat. Assoc.* **1984**, *79*, 575–583.
(4) Shao, J. Linear Model Selection by Cross-Validation. *J. Am. Stat. Assoc.* **1993**, *88*, 486–494.
(5) Xu, Q.-S.; Liang, Y.-Z.; Du, Y.-P. Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. *J. Chemom.* **2004**, *18*, 112–120.
(6) Aqvist, J.; Medina, C.; Samuelsson, J. New Method for Predicting Binding-Affinity in Computer-Aided Drug Design. *Protein Eng.* **1994**, *7*, 385–391.
(7) Martens, H.; Næs, T. *Multivariate Calibration*; Wiley: New York, NY, U.S.A., 1989.
(8) Abdi, H. Partial least square regression (PLS regression). In *Encyclopedia of Measurement and Statistics*; Salkind, N. J., Ed.; Sage Publications, Inc.: 2006; Vol. 2.
(9) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A Comparison of Methods for Modeling Quantitative Structure-Activity Relationships. *J. Med. Chem.* **2004**, *47*, 5541–5554.
(10) Klebe, G.; Abraham, U.; Mietzner, T. Molecular Similarity Indices in a Comparative Analysis (CoMSIA) of Drug Molecules to Correlate and Predict Their Biological Activity. *J. Med. Chem.* **1994**, *37*, 4130–4146.
(11) Visco, D. P., Jr.; Pophale, R. S.; Rintoul, M. D.; Faulon, J.-L. Developing a methodology for an inverse quantitative structure-activity relationship using the signature molecular descriptor. *J. Mol. Graphics Modell.* **2002**, *20*, 429–438.
(12) Free, S. M.; Wilson, J. W. A Mathematical Contribution to Structure-Activity Studies. *J. Med. Chem.* **1964**, *7*, 395–399.
(13) Coats, E. A. The CoMFA Steroids as a Benchmark Dataset for Development of 3D QSAR Methods. *Perspect. Drug Discovery Des.* **1998**, *12−14*, 199–213.
(14) Hastie, R.; Dawes, R. M. *Rational Choice in an Uncertain World: The Psychology of Judgment and Decision Making*; Sage Publications: Thousands Oaks, London, New Delhi, 2001.
(15) Wikipedia. http://www.wikipedia.org (accessed Feb 12, 2008).
(16) Encyclopedia Britannica. http://www.britannica.com/EBchecked/topic/545185/Herbert-A-Simon; http://pubs.acs.org/cgi-bin/abstract.cgi/jcisd8/2006/46/i01/abs/ci049612j.html]. (accessed Feb 19, 2008).
(17) Schölkopf, B.; Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; The MIT Press: Cambridge, MA, 2002.
(18) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.; Sheridan, R.; Feuston, B. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.
(19) Boström, J.; Hogner, A.; Schmitt, S. Do Structurally Similar Ligands Bind in a Similar Fashion. *J. Med. Chem.* **2006**, *49*, 6716–6725.
(20) OEChem Python Library 1.5.1; VIDA 2.1; OpenEye Inc.: Santa Fe, AZ, U.S.A., 2007. http://www.eyesopen.com (accessed Feb 11, 2008).
(21) Peterson, S. D.; Schaal, W.; Karlé, A. Improved CoMFA Modeling by Optimization of Settings. *J. Chem. Inf. Model.* **2004**, *46*, 355–364.
(22) PyMOL 1.0; DeLano Scientific LLC: Palo Alto, CA, U.S.A., 2008. http://pymol.sourceforge.net (accessed Feb 15, 2008).
(23) Mickelson, K. E.; Forsthoefel, J.; Westphal, U. Steroid-Protein Interactions. Human Corticosteroid Binding Globulin: Some Physicochemical Properties and Binding Specificity. *Biochemistry* **1981**, *20*, 6211–6218.
(24) Klieber, M. A.; Underhill, C.; Hammond, G. L.; Muller, Y. A. Corticosteroid-binding Globulin, a Structural Basis for Steroid Transport and Proteinase-triggered Release. *J. Biol. Chem.* **2007**, *282*, 29594–29603.
(25) Tirado-Rives, J.; Jorgensen, W. L. Contribution of Conformer Focusing to the Uncertainty in Predicting Free Energies for Protein-Ligand Binding. *J. Med. Chem.* **2006**, *49*, 5880–5884.
(26) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, NY, U.S.A., 2001.

CI800009U