CAUTION: Popular "Benchmark" Data Sets Do Not Distinguish the Merits of 3D QSAR Methods

John Manchester and Ryszard Czerminski*

AstraZeneca Pharmaceuticals LLP, 35 Gatehouse Drive, Waltham, Massachusetts 02451

Received February 9, 2009

The quality of 3D QSAR models obtained using extremely simple descriptors was examined for nine popular data sets, including the well-known set of 31 steroids, which for 20 years has been the standard for benchmarking 3D QSAR methods. The atomic numbers of atoms coinciding with vertices of the molecular alignment as well as binary descriptors indicating the occupancy of those vertices were compared to models obtained using SAMFA descriptors, which have been shown to yield models statistically indistinguishable from CoMFA. For most data sets, only a minor loss in model performance was observed, even for the occupancy descriptors, where all chemical information is neglected. As a further simplification, models were fitted using descriptors from just a few atomic positions occupied in the majority of active or inactive compounds. No further loss in performance was observed, even though in one case descriptors from a single atomic position were used, and in all cases the number of atomic positions required was fewer than twelve. The resulting models suggest that simply filling space at a few key atomic positions is responsible for enhanced activity. At least for the steroids, this finding is at odds with known SAR and binding interactions with the relevant receptor. Using a simulated data set, we illustrate that this paradoxical outcome is a symptom of having too few observations to describe the response in a data set. It is concluded that none of the nine data sets examined can reliably distinguish the merits of different 3D QSAR descriptors and that they should not be used for this purpose. We advocate the use of simulated data, instead.

1. INTRODUCTION

In a previous paper we explored the idea of using a simplified set of atomic features instead of molecular field descriptors in 3D QSAR.[1] On nine data sets from the literature, models obtained using the atomic descriptors were statistically indistinguishable from those obtained using CoMFA field-based descriptors. Owing to their inherent simplicity and ease of interpretation, SAMFA descriptors were proposed as a useful component of 3D QSAR methodology. However, it was not clear why the simplified descriptors performed so well.

In the present study, we attempted to find just how far the atomic description can be simplified in order to shed light on this question. First, we examined what happens if atom type information is lost from SAMFA by using the atomic numbers of heavy atoms in the 3D superposition as descriptors. We then explored, as the extreme case, what happens if all chemical information is suppressed, using binary descriptors that indicate the occupancy of the various heavy atom positions (vertices) of the molecular alignments. In both cases, we observed only minimal degradation in model performance. The fact that the occupancy descriptors perform so well is counterintuitive but not without precedent. The prerequisite for steric complementarity between a ligand and receptor for binding has long been recognized,[2] and that notion has been used with remarkable success to build models, for example using the MTD method,[3] which correlates differences in steric complementarity to a hypothetical binding pocket to differences in activity among a set of ligands.

It is also plausible that the insensitivity of q^2 to descriptor simplification could be due to statistical artifact. For example, it could be that certain atomic positions are occupied in active compounds, and unoccupied in inactive compounds. In this case, excellent models could be contrived by simply detecting when a few key atomic positions are occupied. Such models would fit the data very well and could exhibit excellent q^2 values. Chemically, however, these models would be useless. Unfortunately, for the nine data sets examined, this turns out to be the case.

For each aligned data set, we isolated a few atomic positions that are selectively populated in active (or inactive) compounds. That is, positions that are occupied in the majority of actives, but missing from the inactives (or vice versa). In all cases, descriptors from 12 or fewer atomic positions produced models indistinguishable from the full descriptor sets. In fact, for the COX-2 data set, descriptors from a single atomic position gave models with q^2 values very similar to those using all CoMFA descriptors. In most cases, no loss in performance was observed, indicating that the full models are indeed supported by a few key atomic positions. These key positions can lie in regions of the compounds that seem to have little to do with the known SAR. Moreover, little or no difference in q^2 was observed between models using atomic number and occupancy descriptors on the key positions. The role of these key positions is not simply to fill hydrophobic voids. More often than not, the key positions are occupied by a polar atom (N or O) in the actives. All nine data sets examined possess

* Corresponding author phone: (781)839-4844; fax: (781)839-4304; e-mail: ryszard.czerminski@astrazeneca.com.

10.1021/ci9000508 CCC: $40.75
Published on Web 05/13/2009

such key atomic positions, which are usually sampled by a single atom type.

One problem with this sort of undersampling is that models with high q^2 values can be fitted using such data sets, irrespective of the (3D) descriptors used. A much more serious problem is that such models could potentially mislead investigators into believing that they are predictive and robust. But those models will not be robust and may fail unexpectedly to give reasonable predictions when used, for example, in virtual screening. Using an external test set to validate the models also will not help, unless careful scrutiny is applied to ensure that the test set presents a variety of chemical features at the positions of interest.

In the ideal case, a data set which has "key positions" should be balanced by including inactive compounds in which those positions are also occupied. Those positions that do make specific interactions with the receptor should be sampled by different atom types in order to challenge the ability of the descriptors to discern the effects of chemical composition on activity.

In practice, however, it is impossible to construct such a data set. Compounds that we would like to include in the data sets cannot be obtained or synthesized, and sometimes existing compounds must be excluded due to uncertainties or difficulties in obtaining the relevant biological data. But there is another reason that makes constructing a balanced data set for any reasonably interesting problem practically impossible. For any system with more than a few key atomic positions (i.e., degrees of freedom), the sheer number of compounds required to sample the biological response surface sufficiently to support a reliable model is prohibitively large. Thus, "real" data sets grossly undersample the response surface, resulting in poor models in which any appearance of quality is due to artifact. We illustrate this point using simulated data. It is therefore concluded that none of the nine data sets examined is suited for distinguishing the merits of different 3D QSAR approaches and that they should not be used for that purpose.

2. METHODS

Data Sets. The steroid benchmark data set[4] and eight data sets (aligned ligands plus data) compiled by Sutherland et al.[5] were used as previously. The latter eight data sets were compiled for the purpose of comparing extant 3D QSAR methods and made available as Supporting Information by the authors for the development and testing of new 3D QSAR methods.

SAMFA. SAMFA descriptors were calculated as previously.[1] Briefly, an irregular template is constructed from the aligned set of molecules by starting with the largest molecule in the data set, then iterating over the remaining molecules and adding atoms to the template that lie beyond a predetermined cutoff from any template atom (1.2 Å in this and the previous study). The template constructed in this way is roughly equivalent to the hypermolecule in the Minimum Steric Difference (MTD) method of Simon et al.,[6−9] and the more precise terminology used by those authors is adopted in this paper. Namely, the hypermolecule is taken as the union of unique vertices among the aligned molecules in a particular data set, within a given resolution (here 1.2 Å). Thus each atom of every molecule in the data set can be mapped to a distinct vertex in the hypermolecule.

Descriptors are calculated for each molecule by mapping onto the hypermolecule the particular atom type at each vertex. The atom types used were C, N, O, S, P, F, Cl, Br, I, and polar H. In addition, atoms were allowed to assume the following "pharmacophoric" identities: X (halogen); aroC (aromatic C), aliC (aliphatic C); HBA (hydrogen bond acceptor); HBD (hydrogen bond donor); and EWG (electron-withdrawing group). The descriptors are represented using a fingerprint, with one bit for each atom type, for each vertex on the template. Using a fingerprint has the advantage that atoms with multiple characteristics can be captured, for example, the fact that an atom is a nitrogen as well as a H-bond donor. After computation of the descriptors, atomic positions with null variance are dropped.

Simplification 1: Atomic Numbers. The approach for constructing SAMFA descriptors was followed except that the set of allowable atom types was confined to the set C, N, O, S, P. F, Cl, Br, I. Only heavy atoms were considered.

Simplification 2: Occupancy. Again, the approach for constructing SAMFA descriptors was used, except that the only property mapped to the template was whether a non-hydrogen atom exists at a particular vertex. For this purpose the IsHeavy() function was used from the OEChem toolkit.[10]

Simplification 3: "Significant" Vertices. For each data set, models using descriptors from a small number of "significant" vertices were fitted. A "significant active" vertex is one that is present in 50% of actives (defined as 1 standard deviation above mean activity), and absent from 50% of inactives (1 σ below mean activity). A "significant inactive" vertex is just the opposite, namely one that is present in more than half the inactives but absent from at least half of the actives.

Model Validation. Monte Carlo cross-validation (MCCV)[11−13] (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 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 $q^2 = 1 - MSE/Var(Y)$, with

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

and

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

where the response variable $Y = pIC50 = -log_{10}(IC50[M])$ (or pK_i for some data sets, as indicated). Data set partitions which resulted in test samples with $Var(Y) = 0$ were

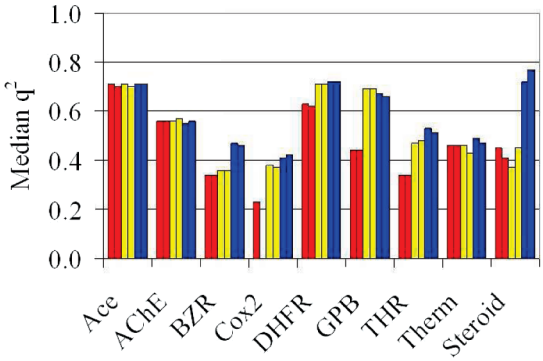

Figure 1. q^2 values obtained for the nine data sets using the three different descriptor sets examined using PLS regression. Each pair of bars represents one descriptor set: occupancy descriptors (red bars); atomic number descriptors (yellow bars); and SAMFA descriptors (blue bars). The first bar in each pair corresponds to the model obtained using all descriptors available for a particular data set; the second bar represents the model obtained using only those descriptors present in at least 50% of active compounds and absent in at least 50% of inactives.

discarded in order to avoid division by zero. For more detailed description see an original SAMFA paper.[1]

3. RESULTS

Key results are summarized here. A detailed analysis of the results for each data set is provided in the Supporting Information. Models were fitted on all nine data sets using Atomic Number and occupancy descriptors and compared to models based on SAMFA descriptors. Figure 1 shows the median q^2 values obtained from all models for each data set. Overall, there is not much difference resulting from simplifying descriptors within each of the data sets, except for the steroids. For the steroids, a large drop in q^2 is observed on simplifying SAMFA descriptors to atomic number descriptors. SAMFA includes descriptors for polar hydrogens, whereas the atomic number descriptors used here are restricted to heavy atoms. Including polar hydrogens with atomic numbers gives q^2 equivalent to SAMFA (data not shown). Similarly, the decrease in q^2 observed for occupancy descriptors is due to their restriction to heavy atoms. The same is true for BZR, the only other data set which shows a slight but noticeable difference between SAMFA and atomic number descriptors.

What is really striking is that there is hardly any difference at all in q^2 between models using descriptors for all vertices and models using just the significant vertices (regardless of the particular descriptors), indicating that all the models are dominated by just a few key atomic positions in the molecular alignment.

In the steroids, four significant vertices were found (out of a total of 44 in the hypermolecule). These are all located within D-ring substituents (Figure 2). Interestingly, each of these vertices is occupied only in active compounds. One explanation for the ability of these four vertices to give reasonable models whether or not specific chemical information is represented by the descriptors could be that they are simply indicating that a key steric void is being filled. This explanation, however, is not supported by the X-ray crystal structure of rat corticosteroid binding globulin (CBG),[14] in which it is shown that the D-ring substituent lies in a conserved hydrophilic region formed by residues Q224,

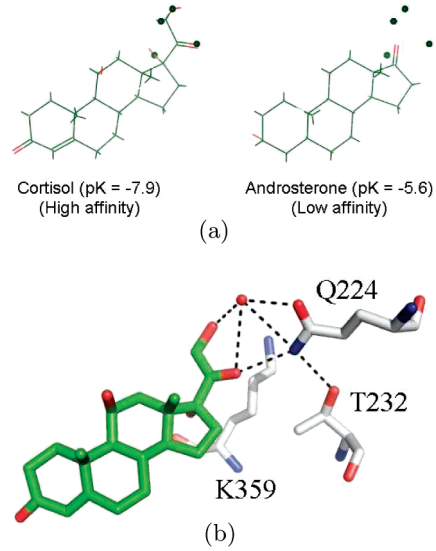

Figure 2. a) The four vertices in the upper right of each image (surrounding the D-ring) are the only 4 out of 44 in the hypermolecule that are present in at least half of active compounds and absent in at least half of inactive compounds. Three of these vertices are only sampled by oxygen (i.e., hydrophilic atoms) in the high-affinity compounds (such as cortisol on the left) and are absent altogether in low-affinity compounds (such as androsterone, on the right). b) The D-ring substituent of cortisol lies in a hydrophilic region in the X-ray crystal structure of rat CBG.[14] Human CBG is conserved in this region, making it unlikely that the occupancy descriptors yield similar models to SAMFA because they represent steric interactions important for binding. Rather, these models highlight the artifactual imbalance in the data set that all compounds with substituents in the D-ring are active. Image created using PyMOL.[15]

T232, K359 (histidine in human), and a bound water. That explanation is also negated by the fact that only one of the vertices is occupied by carbon among active compounds, the other three being occupied only by oxygen. As shown in Figure 2, two of these oxygens make hydrogen bonds with Q224 (one is water-mediated). Based on this observation, changing the nature of these interactions, for example by changing the hydroxymethyl group to an ethyl, would be expected to impact binding affinity. However, such changes are not explored in the data set.

The reason there is no difference between atomic number descriptors and occupancies in the steroid data set is that they amount to the same thing. Particularly for the four vertices which dominate the model, there are no active compounds that populate those vertices with different atom types, and no inactive compounds in which those vertices are populated. The result is a not very useful model which predicts that putting *anything* at those four vertices is good for activity.

ACE, AChE, BZR, and Therm possess essentially the same properties, each dominated by a small number of the total in the hypermolecules: 12 out of 302, 4 out of 153, 3 out of 131, and 13 out of 212, respectively. In each case, the significant vertices were populated either by a single atom type or by different atom types all correlated in the same way with activity.

GPB and THR have five and four significant vertices, respectively. Interestingly, these vertices are populated by different atom types, some of which are correlated with activity, and some of which are anticorrelated with activity.

COX-2 has a single significant vertex which is accessed by different atom types in active and inactive compounds. These data sets all show a decrease in q^2 in going from atomic number to occupancy descriptors (for COX-2, q^2 is obliterated), demonstrating that differences among the descriptors become apparent with more complete sampling of different atom types at each vertex. For DHFR, the four significant vertices found were populated by a variety of atom types. However, these models were dominated by one or two atom types at each vertex, and thus only a slight decrease in q^2 is observed on going from atomic numbers to occupancy descriptors. Although several atom types were sampled at these vertices, improvements in activity correlated with occupancy by one or two atom types are not balanced in the data sets by decreases in activity that are correlated with occupancy by any of the other atom types. The reader is referred to the Supporting Information for a more complete description of the results.

4. DISCUSSION

We have previously shown that CoMFA molecular field descriptors can be simplified to simple atomistic descriptors that consist of crude atom types with no significant loss of q^2 for nine 3D QSAR data sets, including the popular steroid benchmark.[1] In the present paper we have demonstrated that further simplification is possible without drastically affecting q^2. For several of the data sets, the simplification can be taken to an absurd level without affecting q^2 at all, with models based on whether just a few atomic positions are occupied without regard to any chemical information. The most striking observation from this work is that, in all the data sets, it is possible to isolate a few atomic positions that are selectively occupied among the most active compounds, and that these positions are usually only accessed by a single atom type. In this light it is not surprising that the choice of descriptors is relatively unimportant, and that the simple occupancy descriptors perform so well.

The resulting models are dominated by the descriptors corresponding to those key positions. One of the attractive features of 3D QSAR is its potential for uncovering such key positions, which may make important interactions with a receptor. However, if those positions are not sampled by different atoms in compounds with varying activities, then it is impossible to discern whether those positions are indeed important or simply an artifact of the data set. In either case, a model with a reasonable q^2 value can be obtained, which can easily mislead the investigator to assume that the model is useful.

Besides searching for "significant vertices", insensitivity of model quality to descriptor simplification may be a useful diagnostic of artifact in a data set. We illustrate this point using simulated data in the Appendix. We show that the median q^2 obtained falls off more gradually with decreasing sample size (i.e., fraction of the "chemical space" used to build a model) as the simplicity of descriptors increases. That is, simpler descriptors are less sensitive to decreasing sample size than more complex ones, presumably because of the relatively fewer degrees of freedom probed by the simpler descriptors, and consequently the relatively smaller number of compounds needed to support a model of the same quality. For large sample sizes, the limitations of the simpler descriptors are exposed, and the average q^2 obtained increases with increasing descriptor complexity. But at smaller sample sizes, the difference is less pronounced. Below a certain threshold, the complex descriptors perform worse. For SAMFA vs occupancy descriptors on the simulated data set presented in the Apendix, this threshold is about 3%.

In the simulated data set in the Appendix, the hypermolecule consists of four vertices, each of which can possess five possible values (H, C, N, O, or void) for a total of $5^4 - 1$ or 624 possible compounds (the case in which all vertices are void is not used). Because a simple linear scoring function was used to compute the simulated biological activities of these compounds, and all degrees of freedom present in the set are represented by both SAMFA and Atomic+H descriptors, q^2 values of 1.0 are attainable. In fact, a median q^2 of near 1.0 is observed for sample sizes above 10%. As the sample size is further decreased, q^2 rapidly drops off, and below a sample size of about 3%, or 20 compounds, occupancy descriptors perform better than SAMFA on this data set. Thus, when a data set represents less than a certain small fraction of the biologically relevant chemical space, artifactual models are expected.

As an example of a "real" data set, we return to the steroids, of which there are 21 in the original set, and which is sometimes extended to a total of 31 by combining the external test set of 10 (as in the present work). The steroid hypermolecule contains 44 vertices. Allowing only 5 values for each vertex (H, C, N, O, void) equates to about 10^{30} possible compounds, 10% of which is still much more than 31. Even if we exclude the steroid core and consider only heavy atoms, there are still about 10^{17} possibilities.

This analysis assumes that every atomic position is biologically relevant (i.e., equally important to activity) and must thus be included as an explicit degree of freedom. It is extremely unlikely that this assumption is universally valid. In practice, some small number of atomic positions can probably be used to model the ligand−receptor interactions relevant to the measured biological response. A benefit of 3D QSAR is that it can identify such atomic positions without detailed information about the receptor. When this is the case, the performance of simplified descriptors relative to the desired ones may be useful. For example, if occupancy descriptors perform at least as well as more sophisticated descriptors, then it is likely the data set under examination represents an inadequate sample and is likely to produce artifactual models.

Use of an external test set, or the ability of a 3D QSAR model to make predictions, is a less useful diagnostic of model artifact. Verification that the training set adequately samples the chemical space relevant to biological activity is required. Interestingly, a lack of chemical space coverage is usually invoked to justify compounds as "outliers". For example, in the original application of CoMFA, the test-set compound 2-alpha-methyl-9-alpha-fluoro-cortisol (the "notorious steroid 31"),[5] is the only steroid of the 31 examined with a substituent at the 9-position. This compound was poorly predicted by the model, and the authors correctly argued that this was due to poor representation of this chemical feature in the training set. We are arguing here for similar scrutiny of the training set itself. As we have shown (with the benefit of hindsight, not to mention an X-ray crystal

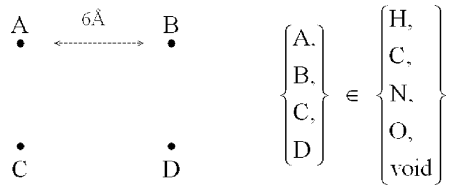

Figure 3. The simulated data set was constructed using "compounds" consisting of 4 nonbonded atoms at vertices separated by 6 Å in the same plane. Each vertex assumes one of five possible identities. "Void" means no atom is present at the corresponding vertex.

structure), without such scrutiny the steroid data set does give rise to artifactual models.

As a testament to their far-reaching vision, the authors of the original CoMFA paper cautioned against the temptation to overinterpret the steric and electrostatic field maps of CoMFA as "receptor maps", since "all possibly relevant aspects of a ligand-receptor interaction surely cannot be explored with test results for a few dozen compounds".[16] We hope that the present work will serve as a complement to 3D QSAR methods by helping practitioners guard against such overinterpretation by looking for a lack of model performance with descriptor simplification, by identifying "significant vertices", or simply by visual inspection to verify that a variety of chemical features are represented at positions of interest among compounds with differing activities. However, even for small data sets, this last approach may be easier said than done, and for large data sets it is probably impossible.

Hopefully, careful attention has been paid to avoid overinterpretation of model results when using the remaining eight data sets examined here. However, all seem to fall into the category that produce artifactual models. All show similar or superior performance for occupancy descriptors relative to descriptors that capture more chemical information. Thus, at least for the purpose of demonstrating differences among 3D QSAR approaches, these data sets are unreliable. Specifically for the purpose of detecting differences among various 3D QSAR approaches, it appears that simulated data are a better choice.

Issues such as sample bias, undersampling, and the "curse of dimensionality" are well-known and often articulated within statistical circles (see, for example, ref 17). They have also long been a source of worry among the QSAR community,[18,19] and attempts to address them have been made. For example, Tropsha has recently stressed the need (developed a diagnostic method) for applicability domain testing,[20] but this approach is designed to assess the robustness of model predictions, not to design or judge the quality of a training set *a priori*. Statistical experimental design approaches have been successfully applied to select robust training sets for QSARs using 2-D descriptors, where a relatively small number of descriptors can be sampled at two or three levels.[21−23] The situation for 3D QSAR, however, is more complicated. The requirement for establishing a molecular alignment based on a (hypothetical) bioactive conformation tends to limit both the size and number of potential training sets, as not all chemical series or compounds within a series are amenable to those requirements. Investigators simply do the best they can with the available data.

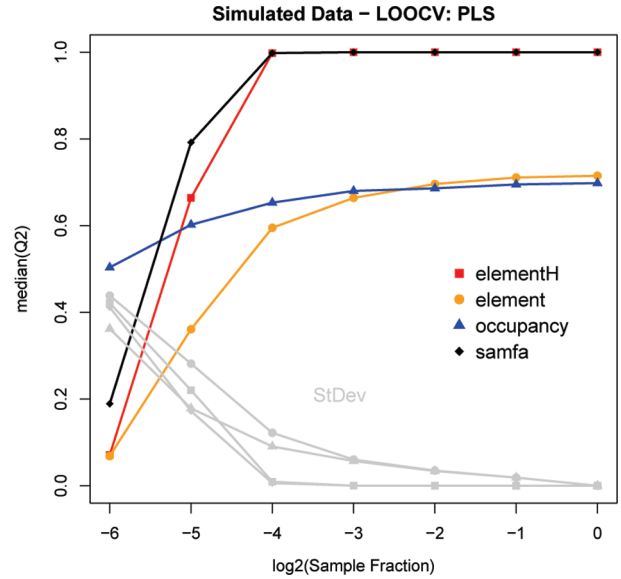

Figure 4. Mean and standard deviation of q^2 as a function of sample size.

From a statistical design standpoint, it is probably impossible in practice to assemble a properly designed data set of real compounds. The great strength of 3D QSAR in describing molecules at a high resolution is also thus its greatest weakness, in introducing more degrees of freedom than can be supported by any reasonable number of compounds.

What are the implications for 3D QSAR? Any set of druglike compounds will produce a hypermolecule with too many vertices to be adequately sampled by a data set of any practical size, according to the above considerations. Of course, not all vertices can participate equally in receptor interactions, and some clues as to which are dominant should be apparent from the SAR, or restrictions of which vertices are interesting may be imposed by synthetic considerations. In either case, the key is to reduce the number of degrees of freedom that require sampling--and then to sample them adequately.

A script is supplied in the Supporting Information which searches an aligned data and lists as output the different atom types represented at each significant vertex found. A second script is included which calculates SAMFA, atomic number, and occupancy descriptors for a data set, which may be used to fit models to determine whether a decrease in q^2 accompanies the decrease in molecular description. We hope that these two scripts will be useful in diagnosing the susceptibility to artifact of other 3D QSAR data sets.

ACKNOWLEDGMENT

The authors gratefully acknowledge the comments of one of the referees of the original SAMFA manuscript, who challenged us to address the question of just how simple the descriptors can be made, and to Dr. Sushmita Lahiri for useful and cheerful discussions.

APPENDIX

A simulated data set was constructed based on "compounds'' that consist of four vertices in a plane, separated by 6 Å from each other, as shown in Figure 3. Each vertex was assigned one of five

values (H,C,N,O,void), giving a total of $5^4 - 1$, or 624 compounds (the compound where all vertices are void was discarded).

"Binding affinities", pK, were computed as the Euclidean distance in eq 1, using the atomic property s, of each compound from an arbitrarily defined "best" compound, in which for this example **A**=C, **B**=H, **C**=N, and **D**=O. The atomic property s was chosen as the product of the Pauling electronegativity and van der Waals radius: H=2.6, C=4.3, N = 4.7 and O=5.1. Voids were given an atomic score of zero. These choices produced values for pK that ranged between 0 and 10, with 10 corresponding to the best compound. SAMFA, atomic numbers for heavy atoms, atomic numbers for all atoms (including H), and occupancy descriptors were then calculated for the entire data set.

$$pK = 10 - \sqrt{\sum_{i}^{vertices} (s_i - best_i)^2} \qquad (1)$$

To simulate "real" data, subsets of varying size were selected at random. Although real data sets are seldom chosen at random, this was considered the least biased and most reasonable approach. For simplicity, geometric sampling rates (based on log_2) were chosen. 256 random subsets were selected for each sample rate, and models were fitted on each with PLS and leave-one-out cross-validation using each set of descriptors.

As expected, the median q^2 increases with increasing sample size (Figure 4). Also as expected, $q^2 = 1$ for large sample sizes for SAMFA and Atomic+H, since both sets of descriptors capture all information present in the simulated data, and PLS is the most appropriate modeling technique here as the simulated binding affinities are constructed by a linear combination of atomic properties. Interestingly, a sample size of between 5% and 10% is needed to realize q^2 values near 1. For sample fractions smaller than 3%, occupancy descriptors actually give higher median q^2 values than SAMFA.

For a small sample fraction (less than about 3%), the variance of q^2 is large, on the order of median q^2. In this region the q^2 obtained is obviously highly dependent on which compounds are included in the sample. These observations are consistent with the notion that the sensitivity to sample size decreases with increasing simplicity of descriptors. We attribute this behavior to the "curse of dimensionality"[17] and the fact that by increasing the complexity of descriptors, the complexity of the design space is effectively enlarged. Thus, it is necessary to use a larger fraction of the whole universe of possible structures to obtain models with reasonable q^2 values.

The minimum sample size required for a 3D QSAR depends on the number of independent degrees of freedom (which parts of the compounds actually affect the biological response and what is the resolution required for representing them) and the shape of the biological response surface (more complex surfaces require more dense sampling). These parameters are difficult, or perhaps impossible, to determine in practice. However, the present work suggests that by assessing the performance of descriptors of varying simplicity, it may be possible to determine whether the sample size is adequate. For example, if occupancy descriptors perform better, then the sample size is certainly too small.

Supporting Information Available: Results for individual data sets and scripts. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Manchester, J.; Czerminski, R. SAMFA: Simplifying Molecular Description for 3D-QSAR. *J. Chem. Inf. Model.* **2008**, *48*, 1167–1173.

(2) Fischer, E. Einfluss der Configuration auf die Wirkung der Enzyme. *Ber. Dt. Chem. Ges.* **1894**, *27*, 2985–2993.

(3) Simon, Z.; Szabadai, Z. Minimal steric difference parameter and the importance of steric fit for structure-biological activity correlations. *Stud. Biophys.* **1973**, *39*, 123–132.

(4) Coats, A. E. The CoMFA Steroids as a Benchmark Dataset for Development of 3D QSAR Methods. *Perspect. Drug Discovery Des.* **1998**, *12−14*, 199–213.

(5) 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.

(6) Chiriac, Z. S. A.; Motoc, I.; Holban, S.; Ciubotariu, D.; Szabadai, Z. Receptor site mapping. Search strategy of standard for correlations with minimal steric differences. *Stud. Biophys.* **1976**, *55*, 217–226.

(7) Simon, Z.; Chiriac, A.; Holban, S.; Ciubotariu, D.; Mihalas, G. *Minimum Steric Difference. The MTD method for QSAR studies; Chemometrics Research Studies Series*; Research Studies Press Ltd.: Letchworth, 1984.

(8) Oprea, T. I.; Ciubotariu, D.; Sulea, T. I.; Simon, Z. Comparison of the Minimal Steric Difference (MTD) and Comparative Molecular Field Analysis (CoMFA) Methods for Analysis of Binding Steroids to Carrier Proteins. *Quant. Struct.-Act. Relat.* **1993**, *12*, 21–26.

(9) Muresan, S.; Bologa, C.; Chiriac, A.; Jastorff, B.; Kurunczi, L.; Simon, Z. Comparative structure-affinity relations by MTD for binding of cycloadenosine monophosphate derivatives to protein kinase receptors. *Quant. Struct.-Act. Relat.* **1994**, *13*, 242–248.

(10) OpenEye Inc. OEChem Python Library 1.5.1; VIDA 2.1, 2007. http://www.eyesopen.com (accessed Feb 11, 2008).

(11) Picard, R. R.; Cook, R. D. Cross-Validation of Regression Models. *J. Am. Stat. Assoc.* **1984**, *79*, 575–583.

(12) Shao, J. Linear Model Selection by Cross-Validation. *J. Am. Stat. Assoc.* **1993**, *88*, 486–494.

(13) 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.

(14) 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.

(15) DeLano Scientific LLC. Palo Alto, CA, U.S.A., PyMOL 1.0, 2008. http://pymol.sourceforge.net (accessed Feb 15, 2008).

(16) 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.

(17) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, NY, U.S.A., 2001.

(18) Martin, Y. C. 3D QSAR: Current State, Scope and Limitations. *Perspect. Drug Discovery Des.* **1998**, *12/13/14*, 3–23.

(19) Oprea, T. I. 3D QSAR modeling in drug design. *Comput. Med. Chem. Drug Discovery* **2004**, *571*, 571–616.

(20) *Chemoinformatics Approaches to Virtual Screening*; Varnek, A., Tropsha, A., Eds.; Royal Society of Chemistry: 2008.

(21) Linusson, A.; Gottfries, J.; Olsson, T.; Ornskov, E.; Folestad, S.; Norden, B.; Wold, S. Statistal Molecular Design, Parallel Synthesis and Biological Evaluation of a Library of Thrombin Inhibitors. *J. Med. Chem.* **2001**, *44*, 3424–3439.

(22) Olsson, I.-M.; Gottfries, J.; Wold, S. D-optimal onion designs in statistical molecular design. *Chemom. Intell. Lab. Syst.* **2004**, *73*, 37–46.

(23) Norinder, U.; Hogberg, T. Quantitative structure-activity relationships and experimental design. In *Textbook of Drug Design and Discovery,* 3rd ed.; Krogsgaard-Larsen, P., Liliefors, T., Madsen, U., Eds.; 2002; pp 127−170.

(24) Bichard, C. J. F.; Mitchell, E. P.; Wormald, M. R.; Watson, K. A.; Johnson, L. N.; Zographos, S. E.; Koutra, D. D.; Oikonomakos, N. G.; Fleet, G. W. J. Potent inhibition of glyocgen phosphorylase by a spirohydantoin of glucopyranose: First pyranose analogues of hydantocidin. *Tetrahedron Lett.* **1995**, *36*, 2145–2148.

(25) Martin, J. L.; Johnson, L. N.; Withers, S. G. Comparison of the binding of glucose and glucose 1-phosphate derivatives to T-state glycogen phosphorylase b. *Biochemistry* **1990**, *29*10745–10757.

(26) Melville, J. L.; Hirst, J. D. TMACC: Interpretable correlation descriptors for quantitative structure-activity relationships. *J. Chem. Inf. Model.* **2007**, *47*, 626–634.

CI9000508