Journal of Molecular Graphics and Modelling 27 (2008) 225–232

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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

Fingerprint-based clustering applied to define a QSAR model use radius

D.G. Sprous *

Redpoint Bio Inc., 7 Graphics Drive, Ewing, NJ 08628, USA

ARTICLE INFO

Article history:
Received 4 December 2007
Received in revised form 4 April 2008
Accepted 24 April 2008
Available online 3 May 2008

Keywords:
QSAR validation
GRAS
Kinase inhibitors
Skin permeability

ABSTRACT

In ongoing research, QSAR has been a tool applied to evaluate compound qualities associated with skin permeability and membership in either a druglike class or specific nondruglike type classes. A need that arose from this pursuit was to know the boundaries of the QSAR models within which molecules could be analyzed. To satisfy this need, a method of QSAR model validation was developed which moves away from the simple declaration of correlation to a description of expected correlation as a function of similarity to the training set. This extension of the "validation" and "predictive" concepts to include a border is referred to henceforth as the *QSAR model use radius*. By defining this metric, it is possible to select for models which have predictivity exterior to their training sets. The heart of this approach is the common use of division into training sets and test sets to demonstrate an ability to successfully predict outside of the training set. The new rigor introduced is to repetitively cluster and systematically increase the permitted dissimilarity within those clusters. The training sets are assembled by taking one and only one compound from each cluster at a specific level of permitted dissimilarity. The QSAR model is developed over these training sets and applied to predict the remaining compounds. In this manner, it is possible to point where there is adequate similarity to predict a compound and where there is not. This method is especially useful for large, chemically redundant systems of greater than 250 compounds where leave-one-out crossvalidation is of limited use. To illustrate this technique, the results of defining the use radius for (a) a skin permeability model (based on 276 compounds), (b) a drug compound and "safe" compound partition (3000 compounds) and (c) a kinase inhibitor and drug compound partition (~1300 compounds) are discussed.

1. Introduction

Medicinal chemistry has a fundamental principle that similar structures can be expected to have similar biological activities. This implies that when descriptors are calculated on such compounds, that the compounds will have coordinates in this descriptor space that define a common volume that is different than compounds that are divergent in biological activity. This volume will not necessarily be regular or continuous, but it will logically be distinct as compared to volumes defined by compounds divergent in biological activity. It is possible to imagine and visualize that QSAR models (or a genetic algorithm or a neural network or other training set-dependent method), trained over a particular set of compounds will be applicable to neighboring molecules but not to molecules farther from the training set. This mental visualization can lead to the understanding that a QSAR model has predictivity (expressible as any number of valid statistics that include Pearson's R^2, q^2, accuracy or the like) but that predictivity is expected within a boundary. This boundary becomes the use radius for a QSAR model where a certain predictivity can be demonstrated and expected.

QSAR models are optimized by computational methods under supervision of a practioner who can control several variables including nature of training set, number of components, descriptors used and so forth. If the only success criterion for a QSAR model development is performance within the training set used in the optimization process, it is in hindsight unsurprising that the Kubinyi Paradox – that the "best" QSAR models have limited predictivity exterior to their training sets – has been a significant problem [1,2]. Causes for poor performance include both extrapolation and overfitting. Overfitting, the act of training a model such that it can accurately predict the training set but has no ability to predict outside the training set, has long been recognized as a problem [1–3]. Extrapolation beyond the applicability or use radius has been more neglected, but recent a online review [4] and a planned ACS symposium for spring 2008 seem to signify that certain frustrations are pressing the issue to the forefront.

The author has frequently been concerned as to demonstrate that a model has value and can predict exterior to the training set. This desire appears in two distinct stages: first, when the model is

* Tel.: +1 609 637 9700; fax: +1 609 637 0126.
E-mail address: dennis.sprous@redpointbio.com.

1093-3263/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2008.04.009

being developed, the need is to optimize the selection of QSAR settings (descriptors choose, dataset trained over, number of components used, ...) to predict exterior to the training set; and second, after the model is in use, it is desirable to know whether a given unknown molecule can be predicted accurately with the QSAR model. These desires led in turn to a method, presented here, which helps to define the performance expectation for a QSAR model against molecules of some known similarity to its training set and second, a given new compound against a QSAR model. The purpose of this paper is to present this method, based on clustering with MACCS structural fingerprints as implemented in MOE [5], that allow the practitioner to develop easy to comprehend metrics of how model performance degrades as diversity from the training set increases. This, in turn, allows metrics to evaluate how confidently a particular molecule can be evaluated with a particular QSAR model.

The exact opposite of this approach is to develop a model and to consider only the statistics of the model over the training set. Though this cannot be considered wise, such studies do commonly appear in peer reviewed publications (a short list of such is can be found in Golbraikh and Tropsha [6]). In the cases where training sets are below a dozen compounds, this is an perhaps an unfortunate necessity. For these small datasets, each compound with associated descriptors and activity is possibly a crucial required point for building the model. In case where the training set exceeds a dozen, there is no logical reason to not at least attempt and report a leave-one-out (LOO) crossvalidation.

LOO crossvalidation is a good first step, and for datasets on the order of 12–30 possibly the best option for demonstrating QSAR predictivity. However, for datasets greater than 30, some consideration should be devoted to challenging the QSAR model method with a reserve test sets or to validate the model by other means. In this scenario, a portion of the dataset is reserved and not used to train the QSAR model. Thereafter, the QSAR model is applied to predict the values of the reserve test set and this is reported as a measure of predictivity. This is necessary since as datasets grow, the chances of a given molecule having effectively a close sibling molecule which is highly comparable in terms of simple activity and in terms of descriptor values increases [7]. Under the circumstance that each molecule in a dataset has such one or more close sibling molecules, LOO crossvalidation becomes useless. In these cases, the omission of a molecule from training is not actually done since a near duplicate is still used in developing the QSAR model. The classical symptom of this is when the LOO-crossvalidated statistics are essentially identical to the raw unvalidated statistics of the QSAR model. This is a guaranteed situation for datasets of greater than 50 compounds.

Clark [7] was cognizant of these issues when he developed a method called progressive scrambling (available in the SYBYL suite of programs) to be applied to moderately large and redundant datasets. In Clark's [7] work at the time, redundant but relatively small datasets were employed that were less than 50 members. The fundamental assumption behind this approach is that perturbing the activities should lead to proportional loss of model robustness as seen by q^2 for a generated QSAR model. In this method, the compounds are sorted according to activity and placed in bins. Within each bin, the activity was swapped between bin members and a QSAR model created. The process was repeated for progressively larger bins. At each stage, q^2 was calculated along with the Pearson's correlation coefficient R^2 between the scrambled activities and the true activities, the latter metric dubbed ryy'^2. A quadratic fit was performed relating q^2 as a function of ryy'^2 and the first derivative calculated for q^2 with respect to ryy'^2. These metrics permitted a concrete measurement of the degree that QSAR models were degraded as larger bins led to larger disparity between descriptors and activities. QSAR models based on descriptors with a real relation to the activities would show gradual degradation while QSAR models that were exercises in overfitting would show more dramatic changes in q^2 as a function of ryy'^2. The process could be used to develop arguments for the correct number of components to use for critical QSAR models where predictivity was crucial.

Golbraikh and Tropsha [6] developed a systematic means of creating partitions between test and training set. Further, they noted that training and test sets should span the same chemical space to be a valid test. By application of clustering protocols, they developed and illustrated methods for creating test and training sets which satisfied the requirement of spanning the same chemical space. Additional statistics were introduced to define when a QSAR model was predictive. The authors used a series of $<$100 member datasets to demonstrate the techniques and illustrate how inspecting only q^2 or R^2 or a crossvalidated metric would be inferior to application of well-constructed training and test sets.

However, neither the methods of Clark et al. [7], Golbraikh and Tropsha [6] or others have systematically connected that predictivity needs to be – when possible – defined within a boundary. Only when this boundary is defined can a new compound be known to be within or exterior to the QSAR model's use radius. This use radius is not a true/false concept but rather an expected performance for given conditions. Compounds closer to the training set will have better expectation to be accurately predicted while compounds farther out will have lower expectations. The present paper addresses this use radius concept with simple series of graphs which show how QSAR model predictivity degrades with respect to diversity between training and test sets. By simple inspection of the curve, it is possible to make statements concerning the expected predictive accuracy for a new compound within some similarity metric to the training set. The graphs can themselves be used to compare different QSAR models where, for instance, descriptors employed of number of components employed were varied, and make judgments as to which choices produce a superior model. The paper illustrates these concepts using two different QSAR techniques, traditional PLS and binary QSAR. This paper uses datasets which are publicly available. The datasets are each at least 276 compounds and are characterized by structural redundancy and imperfect activity data which may come from different sources. The method employed herein is offered as a means to judge how far a QSAR model can be applied and to relate similarity of new compounds to training set to expected QSAR model accuracy for those specific compounds. The author considers this approach – connecting expected model performance to similarity of new compounds to training set compounds – to be a critical additional method for QSAR models that will actually be used.

2. Materials and methods

2.1. Datasets

The datasets employed are listed in Table 1. All these datasets have been used in prior work and the original papers discuss the rationale for the descriptors sets employed [8–10]. The author believes that most readers will comprehend the idea of a "potent kinase inhibitor", a "marketed pharmaceutical", or a "skin absorption flux" dataset but will not be familiar with the GRAS dataset. The GRAS dataset is comprised of compounds from the FEMA/RIFM affirmed list of Generally Recognized As Safe compounds [11–13]. These compounds are permitted as flavorants and are considered to have no biological activity at the quantities

Table 1
Datasets

Name	Shorthand	Count	Description	Ref.
GRAS	GRS	1881	A subset of the FEMA/RIFM affirmed Generally Recognized As Safe flavorants	[9]
Prestwick	PRS	1149	A HTS screening dataset comprised of off patent pharmaceuticals	[19]
Kinase inhibitors dataset	KID	258	A list of potent kinase inhibitors obtained from literature	[8]
Magnusson skin flux	MSF	276	A set of compounds with known measured skin permeability flux	[10]

used in foods. These are not required by regulation to be natural products but largely are. The combined pressures for *safety* and *taste* make these compounds distinct from pharmaceuticals [9]. This dataset was assembled starting from the Flavor-Base database [14] of compound names and converted to SMILES using the Lexichem program [15].

The kinase inhibitor database (KID) consists of 258 potent kinase inhibitors drawn from three reviews [16–18] and previously used to develop a kinase inhibitor recognition model [8]. The inhibitors were not focused on any one specific kinase group and the assays used to determine potency were naturally not uniform. Hence, this dataset is simply a membership list which as a set can help define kinase inhibitor likeness. The author did exactly this previously, developing a kinase inhibitor recognition function [8]. This function was used successfully as part of a HTS kinase focus library purchase decision. KID was previously assembled by manual entry by the author, Zhang and Wang (formerly of CytRx Labs, help gratefully recognized).

The Prestwick database (PRS) is sold as a HTS library [19]. The library consists of 1120 off patent pharmaceuticals and offers a convenient means of representing pharmaceutical ingredient space.

The Magnusson skin flux dataset is 276 compounds with associated known values for skin permeability flux [10]. This is available as a supplemental dataset of names [10]. Again, as was done with the GRAS set, the names were converted to SMILES strings with the program Lexichem [15]. The Magnusson dataset spans from −12.7 to −3.59 over the logJmaxb (log of the skin permeability flux).

Working sets modeled (Table 2) were the union of the GRAS and Prestwick, the union of Kinase and the Prestwick and the Magnusson. The first two working sets were approached as a membership problem or a identify "actives" from "inactives" type problem. In this, we ask can the two original datasets be recognized by a model. This type of problem is well approached with Labute's Binary QSAR method [20] which we employed. The Magnusson working set by contrast had one membership roster and measured skin permeability flux values over several orders of magnitude. This was approached as an effort accurately predict known, measured rates for compounds to pass across the skin (skin flux). This is a problem traditionally and best approached by a method such as partial least squares (PLS).

2.2. Protocols

Most computations described in this paper were done in the MOE suite [5]. Key modules employed include the *Database* modules *Wash*, *Fingerprints*, *Descriptors*, ..., *QSAR-Model*, *Cluster-Codes* and *Model-Evaluate*. In addition, the program *Lexichem* [15] proved invaluable in database generation, specifically where noted in Table 1. Lastly, data processing and graph generation was developed in Microsoft Excel [21] and GraphPad Prism 5.0 [22].

All databases were stored as a MOE Molecular Database (MOE-MDB). The datasets were first ran through the *Wash* utility. Options were set to charge all bases and acids, add hydrogens and remove spurious counter ions. MACCS Fingerprints were generated for each entry of each dataset. Descriptors were calculated over each dataset using the *Descriptor* module and in house developed SVL scripts. Descriptors employed with each working set are listed in Tables 3–5 respectively for GRAS:PRS, KID:PRS and MSF. The descriptor selection rationale for GRAS:PRS has been reported previously [9]. The Sprous et al. paper [8] describes the descriptor selection rationale for a dataset very similar to KID:PRS and was taken as template. The original Magnusson et al. paper [10] provided a clear conceptual template for descriptor set 1 in Table 5. In the MSF set, the present author substituted a_heavy (number of heavy atoms) for molecular mass-based. The a_heavy captures true size-dependent behavior without inappropriate penalties associated with, for instance, an oxygen to sulfur change. The set 2 descriptors were devised to show that descriptors do in fact matter and not just any set of descriptors can be applied to the MSF dataset.

Clustering, *QSAR model generation* and *QSAR model validation*. The *Cluster-Codes* module of MOE [5] was used to generate clusters based on the previously generated MACCS Fingerprints. The tolerance was systematically varied between 0.85 and 0.25 in Tanimoto correlation coefficient (TCC) space. The higher values naturally led to more numerous clusters with fewer entries per cluster. A single representative from each cluster at a specific value of the Tanimoto correlation coefficient was placed in a training set. The rest was placed in the test set. QSAR models were developed over the training sets and applied to predict the test sets. All variables discussed are presented in relation to how they vary as a function of the TCC used in clustering. By example, the paper will present Accuracy or R^2 for the predicted activities for the test set as function of TCC. The behavior the clustering protocol and details of the predictivity of the QSAR models is discussed below for each of the three working sets listed in Table 3.

A common PLS-based QSAR model was developed for the MSF working set. This multiple linear regression (MLR) variant method is common in computational chemistry and interested readers can consult reference text [23].

The binary QSAR method [20] is substantially less familiar and was employed for GRAS:PRS and KID:PRS and is discussed below. Binary QSAR is not multiple linear regression or partial least squares. Specifically, it does not provide a traditional sum-based

Table 2
Working model sets

ID	Modeled dataset	Method	Dependent variable
GRAS:PRS	GRAS and PRS (Preswick) databases	Binary QSAR	+1 for GRAS membership, 0 otherwise
KID:PRS	KID and PRS databases	Binary QSAR	+1 for membership in the Kinase inhibitor database, 0 otherwise
MSF	Magnusson skin flux database	PLS	log[Jmaxb]—log of skin flux permiability

Table 3
GRAS:PRS descriptors

Name	Description	MOE tag	SMILES
MW	Molecular weight	Weight	
Flexibility	Rotable bonds normalized by total number of bonds	b_1rotN	
log(P)	Labute log(P) model	logP(o/w)	
log(S)	Labute solubility model	logS	
Acceptors	H-bond acceptor count	a_acc	
Donors	H-bond donor count	a_don	
Acidic atoms	Acidic atom count	a_acid	
Basic atoms	Basic atom count	a_base	
Halogen	Halogen atom count		F, Cl, Br or I
Aromatic atoms	Aromatic atom count		a

QSAR equation and coefficients. Binary QSAR works off a true/false criterion rather than a series of activities spanning several log units. Thus, binary QSAR cannot be described by usual metrics such as R^2 or q^2. The critical statistic for binary QSAR is accuracy (number correctly assigned/total number of members).

Binary QSAR was initially developed in the MOE [5] context for modeling high throughput screening (HTS) data that is binary true/false classification of tested compounds for active. This situation is essentially identical to the present GRAS/drug discrimination problem. The fundamental equation for binary QSAR was presented by Labute [20] and is shown below:

$$p(x) = \frac{1}{\left[1 + ((m(I)+1)/(m(A)+1)) \prod_{j=1}^{\text{Nbins}} F(x,j,I)/F(x,j,A)\right]} \quad (1)$$

$p(x)$ is the final reported probability that molecule x is active. This is dependent on the distributions for *active* ($F(x, j, A)$) and *inactive* ($F(x, j, I)$) molecules with respect to specific descriptors. The distribution function employed is

$$F = 0.5 \sum_{k=1}^{\text{Nbins}} \frac{P(k) + 1/c}{c + \text{Nbins}/c} [g(k) - g(k-1)] \quad (2)$$

which is depended on the following terms:

$$g(k) = \text{erf}\left(\frac{P(k) - z}{\sigma\sqrt{2}}\right) \quad (3)$$

$P(k)$ is the population at some specific bin k, σ is the variance of the distribution across the specific descriptor of interest, z is the total population across all Nbins and c is a constant that is solved as part of the regression process. In Eq. (1), the multiplication of the product term by the ratio of actives ($m(A) + 1$) and inactives ($m(I) + 1$) in the training set over which the model is developed ensures that the result ranges between 0 and 1. In the case of a descriptor having a exclusion zone where no population is seen, the product-based model used in binary QSAR will capture the situation. In contrast, a sum of individual terms (as in PLS or MLR) would not treat a exclusion zone well. Typically in linear regression models (whether PLS or LLS), a final equation is presented showing the solved coefficients. For Eq. (1), the equivalent would be a large table of individual populations per bin per specific descriptor, with the number of bins themselves changing from descriptor to descriptor. This particular report would be difficult to digest but the visualization is easily accomplished by use of distribution plots (see Sprous and Salemme [9] for a detailed presentation of distribution plots that show why GRAS and PRS are recognizable from one another). For a binary QSAR, the final model statistic is accuracy—the fraction of compounds correctly identified as either active or inactive.

Table 4
KID:PRS descriptors

Name	Description	MOE tag
Aromatics	Aromatic atom count	a_aro
Atoms	Atom count	a_count
Flexibility	Rotable bonds normalized by total number of bonds	b_1rotN
Heavies	Heavy atom count	a_heavy
F	F atom count	a_F
Cl	Cl atom count	a_Cl
Br	Br atom count	a_Br
I	I atom count	a_I
N	N atom count	a_N
O	O atom count	a_O
S	S atom count	a_S
log(S)	Labute solubility model	logS
Acceptors	H-bond acceptor count	a_acc
Donors	H-bond donor count	a_don
Acidic atoms	Acidic atom count	a_acid
Basic atoms	Basic atom count	a_base
log(P)	Labute log(P) model	logP(o/w)

Table 5
MSF descriptors

Name	Description	MOE tag	Set #
Heavies	Heavy atom count	a_heavy	1
log(S)	Labute solubility model	logS	1
SlogP_VSA0	Surface area "owned" with SlogP weight −10 to −0.40	SlogP_VSA0	2
SlogP_VSA1	−0.40, −0.20	SlogP_VSA1	2
SlogP_VSA2	−0.20, 0.00	SlogP_VSA2	2
SlogP_VSA3	0.00, 0.10	SlogP_VSA3	2
SlogP_VSA4	0.10, 0.15	SlogP_VSA4	2
SlogP_VSA5	0.15, 0.20	SlogP_VSA5	2
SlogP_VSA6	0.20, 0.25	SlogP_VSA6	2
SlogP_VSA7	0.25, 0.30	SlogP_VSA7	2
SlogP_VSA8	0.30, 0.40	SlogP_VSA8	2
SlogP_VSA9	0.40, 10.0	SlogP_VSA9	2

3. Results

3.1. Model development and establishing QSAR use radius

The essential details and behavior of the GRAS:PRS QSAR model has been described previously [9]. In that paper, a random partition was performed to create test and training sets. In this paper, we focus on developing test and training sets using a clustering approach. The compounds of the GRAS:PRS working set were clustered according to MACCS Fingerprints at progressively lower values of the Tanimoto correlation coefficient. This creates increasingly fewer, larger, more diverse clusters. A single member from each cluster is taken to form the training set while all other compound are placed in the reserve set. This process is illustrated graphically in Fig. 1, where one can see training set population increasing with Tanimoto correlation coefficient while test set size decreases. This is not a random or proportional scheme. In Fig. 2, the inherently less diverse GRAS compound set (i.e., the actives) are more likely to end up in the test set as Tanimoto correlation coefficient increases since the GRAS set is less diverse than the PRS set.

The training sets were used to generate binary QSAR models using the binary QSAR method. These QSAR models were saved and used to predict the reserve test sets. The overall accuracy for the prediction the test set for GRAS:PRS is shown in Fig. 3 as a function of Tanimoto correlation coefficient. At a Tanimoto correlation coefficient value of 0.85, the accuracy is 0.79. As the line is followed leftward (smaller training set, larger test set, more compounds per cluster, larger more diverse clusters), accuracy shows little change until past 0.45 Tanimoto correlation coefficient where accuracy

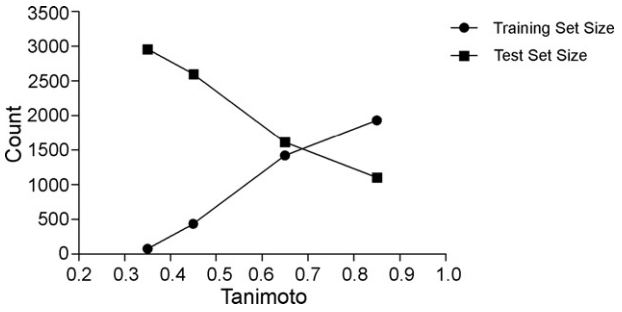

Fig. 1. Population size [count] for the training (circles) and test (squares) sets as function of Tanimoto [Tanimoto] correlation coefficient employed to cluster the united GRAS:PRS modeling set.

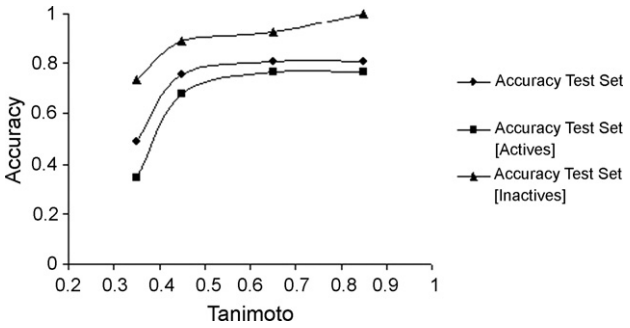

Fig. 3. Accuracy for test set for composite (diamonds), actives (squares) and inactives (triangles) as a function of Tanimoto [Tanimoto] correlation coefficient employed to cluster the united GRAS:PRS modeling set. Actives refers to those members of test set which are GRAS compounds and were correctly predicted by the binary QSAR model to be GRAS compounds. Inactives refers to those members of test set which are from the Prestwick database and were correctly predicted by the binary QSAR model to not be GRAS compounds. Composite is the combined accuracy for both predictions.

drops to 0.75. Thereafter, accuracy starts to show a sharper slope and is chance at 0.35 Tanimoto correlation coefficient (please consider that with a binary QSAR method 0.50 accuracy says that there is no model).

When the composite accuracy is replaced by accuracy for predicting actives and inactives (Fig. 3), it becomes clear that prediction of the actives (GRAS compound) lags behind prediction of the inactives. Still, Figs. 1–5 give us clear evidence that the QSAR model developed over the GRAS:PRS dataset and the descriptors listed in Table 3 shows only slight loss of accuracy as clustering moved from 0.85 to 0.45 Tanimoto correlation coefficient values. Hence, the QSAR model is reported herein to be expected 0.75 accurate overall, 0.67 accurate for actives (GRAS) and 0.90 accurate for inactives. The QSAR use radius that this is true for is 0.45 Tanimoto correlation coefficient, as that this is the point where dramatic changes in the accuracy slope shortly start. Phrased another way, the evidence leads to expect 0.75 accurate overall, 0.68 accurate for actives (GRAS) and 0.89 accurate for inactives within a use radius defined by 0.45 Tanimoto correlation coefficient. At the use radius defined by 0.85 Tanimoto correlation coefficient, the model can be expected to be 0.81 accurate overall, 0.77 for actives (GRAS) and 1.00 accurate for inactives (PRS). The values previously reported for random selection test and training sets were significantly higher than these (more accurate), signifying that the use of a cluster method for test and training sets development is a significantly more demanding quality control protocol.

By review of previous work [9] it can be seen, for this specific GRAS:PRS system, that employment of a FINGERPRINT clustering-based mechanism to generate training and test sets is a more demanding test for the model than that of doing a random partition. A random selection produced no enrichment of GRAS to the test set at the expense of GRAS in the training set. A second observation seen from the previous work is that the accuracy for actives and inactives did not show the same lopsided weight for inactives to be more successfully recognized than actives. In that previous paper, actives and inactive were both hovering near 0.90 accuracy for several trials. This "better" performance tempts the author to conclude that the random selection is the preferred means of creating test and training sets. The divergence in diversity between GRAS and PRS requires a more evenhanded means for developing test and training sets, which the random selection did. In terms of rigor though, it is clear that the FINGERPRINT clustering selection method provides a more conservative lower estimate on accuracy. Further, the random selection provides evidence of predictivity but does not provide an independent means of evaluating where a novel compound would sit relative to the compounds used to develop the model, which by contrast is done in the FINGERPRINT-based clustering method.

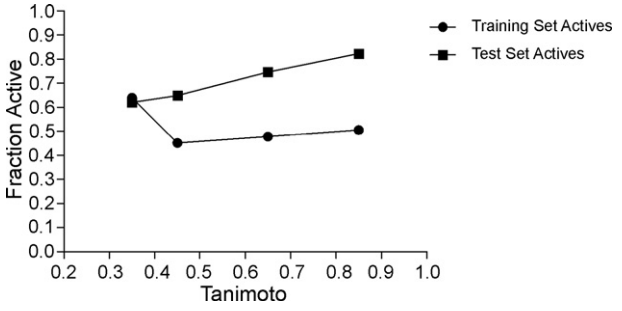

Fig. 2. Fraction active (i.e., member compounds in GRAS dataset) in the training (circles) and test (squares) sets as a function of Tanimoto [Tanimoto] correlation coefficient employed to cluster the united GRAS:PRS modeling set. Note clustering favors actives to be placed into test set at expense of training set.

The KID:PRS working set differs from the GRAS:PRS working set in the substitution of the kinase inhibitors for the GRAS compounds. While the overall sizes of the test and training sets behave the same as was seen with GRAS:PRS (Fig. 4), there are differences in KID diversity that lead to other differences. This can be seen by comparing the behavior seen in Fig. 5 (fractional population of actives in the test and training sets for KID:PRS) to that in Fig. 2 (fractional population of actives in the test and training sets for GRAS:PRS). In Fig. 5, there are random variations as Tanimoto correlation coefficient is varied. These are, compared to Fig. 2, small differences and are certainly not the consistent enrichment of actives to the test set seen in the GRAS:PRS working set. This is interpreted as a signature that KID and PRS are, within MACCS Fingerprints space, essentially the same in terms of diversity, leading to KID occupying a near constant 15–20% of the test and training set with the exception of the smallest training sets seen at the smallest Tanimoto coefficient values. A random selected test and training set method would produce essentially the same profile.

Fig. 6 presents composite accuracy, accuracy for predicting actives and accuracy for predicting inactives. Since the population for actives (KID) is a fifth the population of inactives (PRS), composite accuracy and inactive accuracy are essentially the same. Accuracy over KID (actives) is lower typically. Accuracy on actives for KID:PRS is lower than GRAS:PRS. This may signify that the KID compounds are more similar to the PRS than GRAS, making prediction more difficult. Between 0.85 and 0.65 Tanimoto correlation coefficient, there is little accuracy loss: the inactives and composite are essentially unchanged (accuracy = ~0.94 for

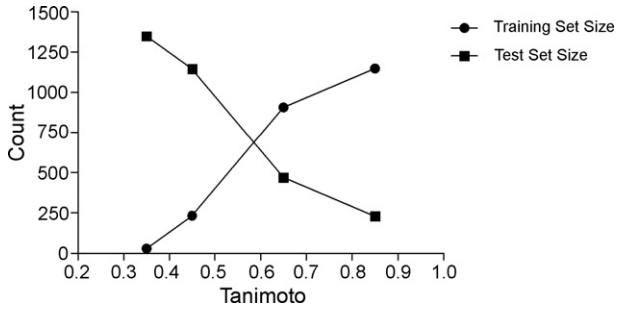

Fig. 4. Population size [count] for the training (circles) and test (squares) sets as function of Tanimoto [Tanimoto] correlation coefficient employed to cluster the united KID:PRS modeling set.

both at both values of Tanimoto correlation coefficient) while the actives drift from 0.84 to 0.77. After 0.65 Tanimoto correlation coefficient, drops of accuracy are more pronounced. The change for the actives (accuracy is 0.66 at 0.45 Tanimoto correlation coefficient) is essentially linear when compared to the values at 0.65 and 0.85 but the change for composite and inactives is steeper. By 0.35, prediction for actives is random and it is clearly moving into a range where diversity is too divergent to predict compounds accurately.

The MSF working dataset is not treated with binary QSAR since this is a problem of predicting log activities, not membership. In this working set, the author modeled the skin permeability flux using two different sets of descriptors (see Table 5). The first set was devised based on published analysis done in Magnusson et al. [10] which showed clearly that the weight and solubility were dominant factors in skin permeability flux determination. Due to the authors experience, he substituted heavy atom count for molecular weight, recognizing that heavy atom count will more accurately capture molecular size phenomena when compounds have mixes of elements from more than a single row. By example, molecular weight considers the change between O and S to be dramatic where heavy atom count does not. The second set of descriptors numbers 10 and is a fractional surface area that contributes to a specified percentage of logP. Regardless of the set of descriptors used, the same training and test sets were employed. In Fig. 7, these set sizes are presented as a function of Tanimoto correlation coefficient. Since this is not a membership question, a figure equivalent to Fig. 2 is not possible.

In all of the figures where R^2 is presented as a function of TCC, it can be tempting to simplify that "more compounds in training set leads to better performance". Naturally, a large training set of several hundred compounds which are no worse than 0.85 Tanimoto similarity to the reserve test set is expected to perform

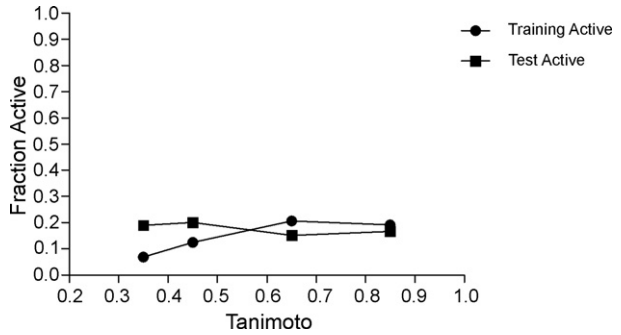

Fig. 5. Fraction active (i.e., member compounds in GRAS dataset) in the training (circles) and test (squares) sets as a function of Tanimoto [Tanimoto] correlation coefficient employed to cluster the united KID:PRS modeling set. Note lack of clustering between test and training sets compared to Fig. 2.

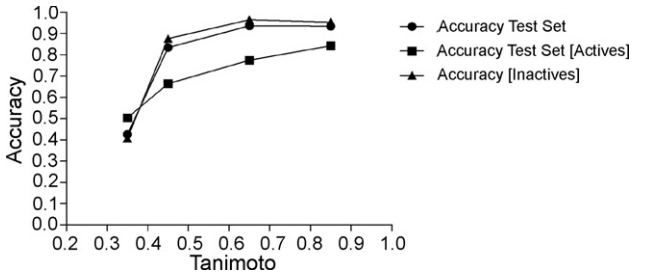

Fig. 6. Accuracy for test set for composite (diamonds), actives (squares) and inactives (triangles) as a function of Tanimoto [Tanimoto] correlation coefficient employed to cluster the united KID:PRS modeling set. Actives refers to those members of test set which are KID compounds and were correctly predicted by the binary QSAR model to be KID compounds. Inactives refers to those members of test set which are from the Prestwick database and were correctly predicted by the binary QSAR model to not be KID compounds. Composite is the combined accuracy for both predictions.

well. More significant in looking at these figures is to consider the slope of R^2 versus TCC. The figures shown here have gradual slopes on the range from 0.40 to 0.85 TCC. This is not accidental, the settings for the QSAR models were optimized to favor this behavior. Alternatively, a higher value for R^2 could have been selected for at 0.85 TCC by varying such parameters as the number of components, descriptor set choice and so forth. Such a model, however, though "better" by a traditional metric is considered here less desirable since it comes at the expense of predictivity farther from the training set. An explicit example of how a QSAR model settings can be selected is given below.

3.1.1. Application: comparing different descriptor sets for model development

Fig. 8 presents the observed Pearson's correlation coefficient as a function for both the training and test set for the model built using the first set of descriptors (logS and heavy atom count). In Figs. 8 and 9, R^2 is similar at 0.85 TCC. Specifically, R^2 for the test set is approximately 0.75 and training set is 0.65 for both models. For the first model based on simply heavy atom count and logS, this situation of higher test set R^2 continues for the range until 0.35 TCC. Thereafter, the R^2 for the training set starts to climb to 1.0 while the R^2 for the test set drops. In Fig. 9, Pearson's correlation coefficient is presented for the second model based on the SlogP-based fractional surface areas. In this model, the R^2 for both training and test mirror the results seen for the first model across Tanimoto correlation coefficient values above 0.60. However, the predictive strength of the model is weaker below 0.60 for the test set than seen in the first model, and the drop in R^2 as a function of Tanimoto correlation coefficient is more rapid and ends at a more dismal predictive point. When all factors are weighed, the fact support that the simpler model is superior in this specific case for

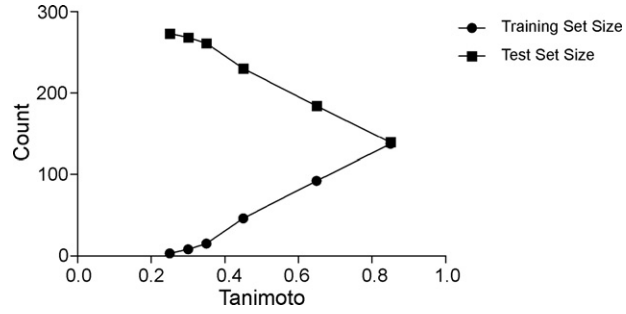

Fig. 7. Training and test set sizes for the MSF dataset as a function of Tanimoto [Tanimoto] correlation coefficient employed to cluster the united MSF dataset.

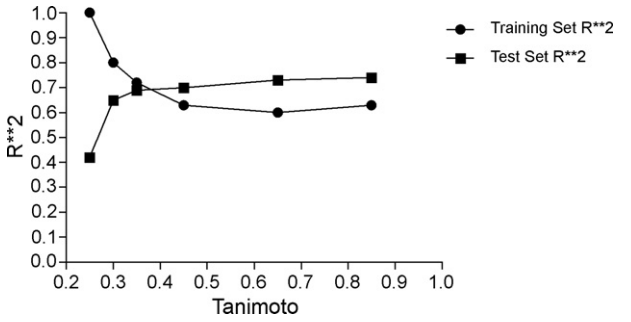

Fig. 8. R^2 (square of the Pearson's correlation coefficient) as a function of Tanimoto [Tanimoto] correlation coefficient employed to cluster the united MSF dataset for Model 1 (the QSAR model based on a_heavy and logS).

predicting skin permeability flux. This result would not be obvious if only the raw unvalidated model was inspected.

3.1.2. Application: should a given compound be predicted by the model?

Table 6 presents several molecules which are either GRAS or pharmaceutical compounds (note column "Description" which clarifies a molecule's origin). Assigned group classifications are presented as per what the GRAS:PRS model determined. These molecules were not in the training set used to develop the model to recognize either GRAS or pharmaceutical compounds. Earlier, it was noted that for this model there was little change in the slope of Accuracy as a function of Tanimoto correlation coefficient between 0.45 and 0.85 TCC, demonstrating a QSAR use radius of 0.45 TCC for this model. As has been noted in the original paper [9], GRAS is a legal definition and does not have chemical requirements but GRAS compounds tend to share common traits produced by a selective pressure for safety and taste.

Compounds CID 22831877 and CID 11739408 are both now presently on the FEMA GRAS approved list. Both compounds are tasteless in themselves but enhance the savory sensation. Compound CID 22831877 has no close analogs within the GRAS database at any value of the Tanimoto correlation coefficient. It is predicted to by a GRAS-like compound, but that result must be considered an extrapolation. The second compound CID 11739408 is predicted to be more pharmaceutical like. This second compound has 573 neighbors at 0.50 TCC and 2711 neighbors at 0.45 TCC. Given this, we must conclude that the result is trustworthy. Again, it is important to understand that this is not a declaration that the compound is not GRAS – which is a legal definition – but that it resembles more pharmaceuticals.

Ethyl cinnamate, CAS: 103-36-6 is a FEMA GRAS list compound. The automated conversion of molecule names to SMILES failed to

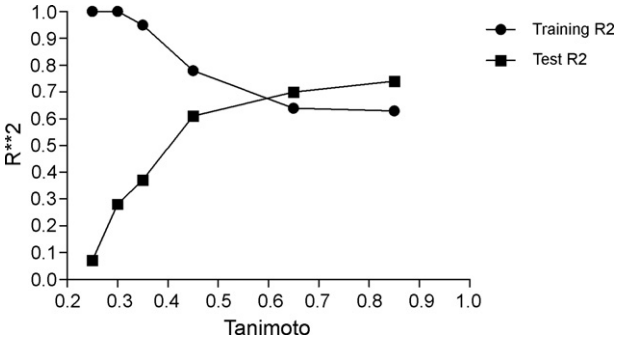

Fig. 9. R^2 (square of the Pearson's correlation coefficient) as a function of Tanimoto [Tanimoto] correlation coefficient employed to cluster the united MSF dataset for Model 2 (the QSAR model based on SlogP_VSAx descriptors).

Table 6
Unknowns and the GRAS:PRS model

ID (CAS or PubChem CID)	Description	GRAS:PRS model assignment	Neighbor population at specific TCC			
			0.45	0.50	0.65	0.85
CID: 22831877	Recent umami enhancer from Senomyx. GRAS	O1c2cc(ccc2OC1)C(=O)NC(CCC)CCC GRAS	0	0	0	0
CID: 11739408	Recent umami enhancer from Senomyx. GRAS	O(C)c1cc(OC)ccc1CNC(=O)C(=O)NCCc1ncccc1 Pharmaceutical	2711	573	0	0
CAS: 103-36-6	Ethyl cinnamate. GRAS	O(CC)C(=O)\C=C\c1ccccc1 GRAS	2711	1180	9	8
CAS: 51115-67-4	O=C(NC)C(C)C)(C(C)C)C Marketed taste coolant. GRAS	GRAS	2	2	2	2
CAS: 220127-57-1	Gleevec–approved pharmaceutical	S(O)(=O)(=O)C.O=C(Nc1cc(Nc2nc(ccn2)-c2cccnc2)c(cc1)C)c1ccc(cc1)CN1CCN(CC1)C Pharmaceutical	2711	0	0	0
CAS: 134523-00-5	Lipitor–approved pharmaceutical	Fc1ccc(cc1)-c1n(CC[C@@H](O)C[C@@H](O)CC(=O)[O-])c(C(C)C)c(C(=O)Nc2ccccc2)c1-c1ccccc1.Fc1ccc(cc1)-c1n(CC[C@@H](O)C[C@@H](O)CC(=O)[O-])c(C(C)C)c(C(=O)Nc2ccccc2)c1-c1ccccc1 Pharmaceutical	2711	573	0	0
CAS: 162011-90-7	Vioxx–approved pharmaceutical	S(=O)(=O)(C)c1ccc(cc1)C=1COC(=O)C=1c1ccccc1 Pharmaceutical	2	2	0	0

return a SMILE for compound CAS: 103-36-6 in our previous paper. This has become fortunate as it can be used as a control. In Table 6, we can see that this compound has a analogs at all levels of the TCC. Given this, the result of the compound be classified as GRAS like is a trustworthy result.

CAS: 51115-67-4 (also called WS-23) is a new GRAS entry which has a strong coolant effect similar to mint. It has two close analogs with which it shares a similarity greater than 0.85 TCC. Given this level of similarity, the result of a GRAS-like classification is trustworthy despite the low number of analogs in the database.

The last three compounds are more modern pharmaceuticals and not present in the Prestwick collection. All are classified as being pharmaceuticals by the model. Gleevec and lipitor have analogs at 0.45 TCC. Gleevec has analogs at 0.50. Given the population of the analogs at those values, the assignment is to be considered trustworthy for both gleevec and lipitor. The situation for vioxx is however, different. It is classified as a pharmaceutical. Vioxx has two compound analogs at 0.45 and 0.50 TCC, strictly speaking making the compound within the use radius of the model. However, given this low population, a certain hesitancy would be called for if no other knowledge was available for the molecule.

4. Summary

The present paper presented a tool for measuring QSAR model predictivity. The method is a means holding descriptors constant but reducing diversity of a training set until to develop an estimate of how distant an unknown compound can be from the training set to have properties predicted with some accuracy. The result is not a single number that can be quoted. Rather, the method creates a graph that represents a measured accuracy for predicting compounds not used to develop the QSAR model as a function of their similarity to compounds used to train the model. Figs. 3, 6, 8 and 9 are representative examples of such graphs. In this manner, it is possible to consider not a simple scalar number as the predictivity of a QSAR model, but rather to consider predictivity within the context of a use radius. For the GRAS:PRS and KID:PRS, we can call the use radius to be 0.45 Tanimoto correlation coefficient since the slope of observed accuracy changes little from 0.45 to 0.85 (see Figs. 3 and 6). For the MSF modeled with the first descriptor set, the use radius is 0.30 Tanimoto correlation coefficient (see Fig. 8). For the MSF modeled with the second descriptor set, the use radius is 0.65, as that the slope changes dramatically after this point. Regardless, the true value of these metrics are to be found by inspection of the graphs, rather than simple quotes of single point numbers. For high value unknown compounds, utilization of this method requires the simple mapping of the distance of the unknown compounds to members of the set used to train the QSAR. In this manner, it is then easy to see an expectation for what predictivity will be for that molecule with a specific QSAR model.

References

[1] H. Kubinyi, F.A. Hamprecht, T. Mietzer, Three dimensional quantitative similarity–activity relationships (3D-QSiAR) from SEAL similarity matrices, J. Med. Chem. 41 (1998) 2553–2564.
[2] H. Kubinyi, Validation and predictivity of QSAR models, in: E. Aki Sener, I. Yalcin (Eds.), QSAR & Molecular Modelling in Rational Design of Bioactive Molecules. Proceedings of the 15th European Symposium on QSAR & Molecular Modelling, Instanbul, Turkey, 2004, CADD Society, Ankara, Turkey, 2006, pp. 30–33.
[3] D.M. Hawkins, The problem of overfitting, J. Chem. Inf. Comput. Sci. 44 (2004) 1–12.
[4] W.A. Warr, Accuracy of prediction, Strand Life Science Pvt. Ltd., http://www.qsar-world.com/qsar-predictivity1.php, 2007.
[5] MOE, MOE 2006.8 Chemical Computing Group, Montreal, Quebec, Canada, 2006.
[6] A. Golbraikh, A. Tropsha, Predicted QSAR modeling based on diversity sampling of experimental datasets for training and test set selection, Mol. Divers. 5 (2002) 231–243.
[7] R.D. Clark, D.G. Sprous, J.M. Leonard, Progressive scrambling: validating models based on large datasets, in: H.-D. Holtje, W. Sipple (Eds.), Rational Approaches to Drug Design, Porous Science, S.A. Barcelona, Spain, 2001, pp. 475–486.
[8] D.G. Sprous, J. Zhang, L. Zhang, Z. Wang, M.A. Tepper, Kinase inhibitor recognition by use of a multivariable QSAR model, J. Mol. Graph. Model. 24 (2005) 278–295.
[9] D.G. Sprous, F.R. Salemme, A comparison of the chemical properties of drugs and FEMA/FDA notified GRAS chemical compounds used in the food industry, Food Chem. Toxicol. 45 (2007) 1419–1427.
[10] B.M. Magnusson, Y.G. Anissimov, S.E. Cross, M.S. Roberts, Molecular size as the main determinant of solute maximum flux across the skin, J. Invest. Dermatol. 122 (2004) 993–999.
[11] G.A. Burdock, The GRAS process, Food Technol. 57 (2003) 17.
[12] J.B. Hallagan, R.L. Hall, FEMA GRAS—a GRAS assessment program for flavor ingredients, Regul. Toxicol. Pharmacol. 21 (1995) 422–430.
[13] I.C. Munro, P. Shubik, R. Hall, Principles of the safety evaluation of flavoring substances, Food Chem. Toxicol. 36 (1998) 529–540.
[14] Flavor-Base, Leffingwell & Associates, Canton, GA, USA, 2004.
[15] LexiChem, OpenEyes, Santa Fe, NM, USA, 2006.
[16] J.L. Adams, D. Lee, Recent progress towards the identification of selective inhibitors of serine/threonine protein kinases, Curr. Opin. Drug Discov. Dev. 2 (1999) 96–109.
[17] A. Bridges, Chemical inhibitors of protein kinases, Chem. Rev. 101 (2001) 2541–2571.
[18] G. McMahon, L. Sun, C. Liang, C. Tang, Protein kinase inhibitors: structural determinants for target specificity, Curr. Opin. Drug Discov. Dev. 1 (1998) 131–146.
[19] Prestwick Chemical Library, Prestwick Chemical Inc., Illkirch, France, 2005.
[20] Labute, Binary QSAR: a new method for the determination of QSARs, Pac. Sym. Biocomp. 4 (1999) 444–455.
[21] Microsoft Excel, Micosoft Inc., Redmond, WA, USA, 2003.
[22] GraphPad Prism 5, GraphPad Software, San Diego, CA, USA, 2007.
[23] A.R. Leach, Molecular Modelling: Principles and Applications, Pearson Education Limited, Essex, England, 1996.