Simple Idea to Generate Fragment and Pharmacophore Descriptors and Their Implications in Chemical Informatics

Cornel Catana*

Drug Discovery Informatics, EMD Serono, 1 Technology Place, Rockland, Massachusetts 02370

Received September 17, 2008

Using a well-defined set of fragments/pharmacophores, a new methodology to calculate fragment/pharmacophore descriptors for any molecule onto which at least one fragment/pharmacophore can be mapped is presented. To each fragment/pharmacophore present in a molecule, we attach a descriptor that is calculated by identifying the molecule's atoms onto which it maps and summing over its constituent atomic descriptors. The attached descriptors are named C-fragment/pharmacophore descriptors, and this methodology can be applied to any descriptors defined at the atomic level, such as the partition coefficient, molar refractivity, electrotopological state, etc. By using this methodology, the same fragment/pharmacophore can be shown to have different values in different molecules resulting in better discrimination power. As we know, fragment and pharmacophore fingerprints have a lot of applications in chemical informatics. This study has attempted to find the impact of replacing the traditional value of "1" in a fingerprint with real numbers derived form C-fragment/pharmacophore descriptors. One way to do this is to assess the utility of C-fragment/pharmacophore descriptors in modeling different end points. Here, we exemplify with data from CYP and hERG. The fact that, in many cases, the obtained models were fairly successful and C-fragment descriptors were ranked among the top ones supports the idea that they play an important role in correlation. When we modeled hERG with C-pharmacophore descriptors, however, the model performances decreased slightly, and we attribute this, mainly to the fact that there is no technique capable of handling multiple instances (states). We hope this will open new research, especially in the emerging field of machine learning. Further research is needed to see the impact of C-fragment/pharmacophore descriptors in similarity/dissimilarity applications.

INTRODUCTION

Methods such as QSAR and machine learning have dramatically increased in popularity over recent years. Advances in theoretical and computational chemistry and more readily available computational power have made this possible. The fact that computational models can sometimes successfully replace some experiments, coupled with increasing pressure on the pharmaceutical industry to become more productive, has provided compelling reasons to expand these fields. The representation of molecules as numerical descriptors has developed from relatively simple forms calculated from two-dimensional (2D) chemical structures to more complex forms representing three-dimensional (3D) chemical structures or complex molecular fingerprints consisting of numerous bit positions representing specific chemical information. Such numerical descriptors are named molecular descriptors, and they can quite often be used to model end points. These end points can be any molecular property or bioactivity that can be experimentally measured. The modeling process is based on a central concept in chemistry that stipulates that properties of a molecule are intimately related to its molecular structure.

In addition to their ability to predict biological activity and ADME/toxicity, these representations are useful in diversity analysis, library design, virtual screening, and other purposes. A large series of molecular descriptors has been published and recently reviewed,[1] but it is not within the scope of this paper to give a comprehensive overview.

Chemical descriptors can be categorized in multiple ways, for example, conformational, electronic, quantum mechanical, topological, spatial, structural, or as 1D, 2D, and 3D descriptors, etc. Here, our focus will be on fragment (substructure) and pharmacophore descriptors. Maximum common substructure (MCS) searches are among the earliest substructure searching algorithms used[2] together with Free-Wilson[3] analysis, which compares the presence or absence of substructural fragments to a biological activity. Other examples of fragment-based descriptors use reduced graphs,[4,5] "molecular tree" fingerprints,[6,7] or related "atom environments".[8−10]

In this study, we have narrowed the area of fragment/pharmacophore descriptors to those that can be calculated as a sum over its constituent atomic descriptors. Well-known atomic descriptors that can be summed at the fragment/pharmacophore level are the E-State (electrotopological state),[11] the partition coefficient (logP),[12] and molar refractivity (MR).[12] How we split a molecule into fragments is not part of this study, but the fragmentation scheme plays an important role and the results depend on it. In terms of pharmacophore descriptors, we concentrated our efforts on graph 3-point pharmacophore fingerprints (which is a 2D pharmacophore), but the methodology can also be extended

* To whom correspondence should be addressed. Phone: (734) 786-3361. E-mail: catana_c@yahoo.com.

10.1021/ci800339p CCC: $40.75
Published on Web 02/25/2009

Table 1. Fingerprint Values

compound	F_1	F_2	F_3	F_4
molecule 1	1	0	1	0
molecule 2	0	1	0	0
molecule 3	0	1	0	0

Table 2. Bits 1 from Table 1 Are Replaced with the Sum of E-State Atom Descriptors over the Atoms Making up That Fragment

compound	CF_EState_1	CF_EState_2	CF_EState_3	CF_EState_4
molecule 1	18.91	0	15.22	0
molecule 2	0	7.39	0	0
molecule 3	0	8.15	0	0

to include 3D pharmacophores. All the work on pharmacophores was done using Molecular Operating Environment (MOE).[13]

METHODOLOGY

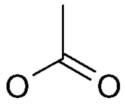
We began by taking a set of molecules and split them into fragments using a certain scheme. Here we used the "Generate Fragments" component from Pipeline Pilot[14] with all the options checked except "MurckoAssemblies" and this treatment is applied throughout the paper if nothing else is mentioned. The number of fragments obtained depends on the scheme used and the size of the data set.

To each molecule we associated a constant length fingerprint (the length depended on the number of fragments generated). When the fragment is present the associated value is 1, otherwise it is 0.

Table 1 is a snapshot of what we obtained for the first 3 molecules and the first 4 fragments from a set of generic molecules split into fragments (F). Each fragment is then numbered (e.g., F_1 stands for first fragment, etc.).

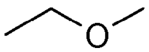
Next, we mapped each fragment back onto the initial molecule and calculated a C-fragment descriptor (in this case based on E-State atom descriptors) by summing over the atoms that constitute the fragment. The values obtained are presented in Table 2 (e.g., CF_EState_1, stands for E-State descriptor from fragment 1; C stands for "comprehensive").

In our case fragment 1 was



For molecule 1, fragment 1, the atom E-State values were: 8.94 for atom number 28, 10.87 for atom number 27, −0.7 for atom number 26, and −0.2 for atom number 24 (see Figure 1). The sum of 18.91 was thus obtained and assigned to CF_EState_1.

In Figure 2, we displayed some of the E-State values for molecule 3; in this case fragment 2 was



Summing over 1.06 for atom number 16, 0.42 for atom number 15, 6.07 for atom number 14, and 0.6 for atom number 13 (see Figure 2), gave a CF_EState_2 value of 8.15 (see Table 2).

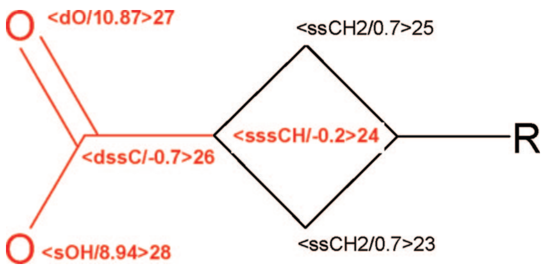


Figure 1. Atom E-State values for fragment 1, molecule 1.

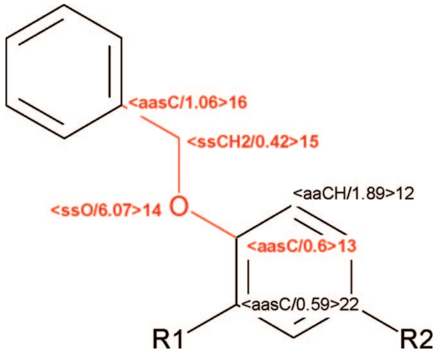


Figure 2. Atom E-State values for fragment 2, molecule 3.

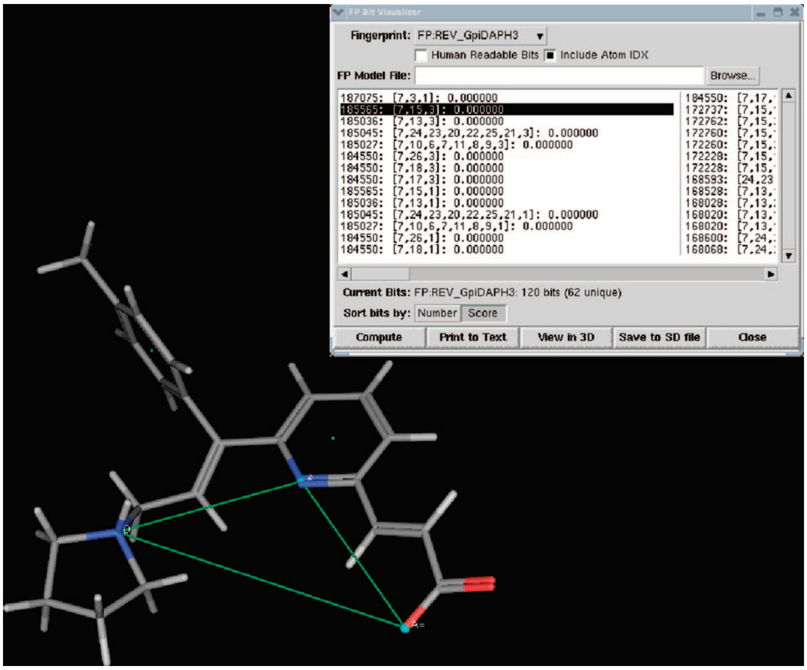
Examination of Tables 1 and 2, shows that "1" from the first table has been replaced with a real number; the same fragment in different molecules has different values (e.g., CF_EState_2) even if the molecules are very similar. We suppose that the discrimination power is increased when "1" is replaced with a real number.

The procedure presented above was extended to other atom descriptors from Pipeline Pilot like AlogP (labeled CF_AlogP_*; "*" substituting the fragment number), MR (labeled CF_MR_*), positive and negative Gasteiger partial charges (labeled CF_GC_P_* and CF_GC_N_*), with the latter being the absolute value of the negative charge. The atomic van der Waals surface areas (VSA; http://www.chemcomp.com/journal/vsadesc.htm) of a fragment on which the partial charge is positive/negative was labeled CF_VSA_P_*/CF_VSA_N_*, respectively; the default VSA for a fragment was labeled CF_VSA_*, and the atomic surface area with AlogP positive/negative was labeled CF_VSA_AlogP_P_*/CF_VSA_AlogP_N_*, respectively. If we had multiple identical fragments in the same molecule, we summed over all of them and reported them as one; other scenarios can be imagined.

In some instances, were we to sum positive and negative values (e.g., GC) the results would be near zero. We wanted to be able to discriminate this from a default zero indicating that the fragment was not present in the molecule. Therefore, we chose to apply absolute values in our calculations.

A Pipeline Pilot protocol to calculate the C-fragment descriptors for all the atomic properties mentioned above has been written. Table 2 exemplifies the output for 3 molecules, 4 fragments, and C-fragment E-State descriptors.

We can build descriptors like those mentioned above not only for fragments (a sequence of connected atoms) but also for atoms that are disconnected such as those involved in pharmacophores. There are several pharmacophore fingerprints (TAD, TGT, TGD, etc.) implemented in MOE,[13] but


FP Bit Visualizer
Fingerprint: FP:REV_GpiDAPH3
☐ Human Readable Bits ☒ Include Atom IDX
FP Model File:
Browse...
Sort bits by: Number Score
Compute
Print to Text
View in 3D
Save to SD file
Close

```
187075:  [7,3,1]:  0.000000                      184550:  [7,17,
185565:  [7,15,3]:  0.000000                     172737:  [7,15,
185036:  [7,13,3]:  0.000000                     172762:  [7,15,
185045:  [7,24,23,20,22,25,21,3]:  0.000000      172760:  [7,15,
185027:  [7,10,6,7,11,8,9,3]:  0.000000          172260:  [7,15,
184550:  [7,26,3]:  0.000000                     172228:  [7,15,
184550:  [7,18,3]:  0.000000                     172228:  [7,15,
184550:  [7,17,3]:  0.000000                     168593:  [24,23
185565:  [7,15,1]:  0.000000                     168528:  [7,13,
185036:  [7,13,1]:  0.000000                     168028:  [7,13,
185045:  [7,24,23,20,22,25,21,1]:  0.000000      168020:  [7,13,
185027:  [7,10,6,7,11,8,9,1]:  0.000000          168020:  [7,13,
184550:  [7,26,1]:  0.000000                     168600:  [7,24,
184550:  [7,18,1]:  0.000000                     168068:  [7,24,
```

Current Bits: FP:REV_GpiDAPH3: 120 bits (62 unique)

Figure 3. Pharmacophore atoms corresponding to bit 185 565 in GpiDAH3 are connected through green lines. Highlighted in FP Bit Visualizer box is the atom index [7,15,3] onto which bit 185 565 maps.

here we exemplify with GpiDAPH3 (graph 3-point pharmacophore which is a 2D pharmacophore) because in our projects it provided good results. These fingerprints have a fixed length; they do not depend on the size of the data set but rather each fingerprint has a predefined size (e.g., GpiDAPH3 has 262,144 bits).

The pharmacophore atoms are mapped to the atom index using a script provided to us by MOE developers. In Figure 3, we represented one (bit number 185 565) of the 262 144 bits of GpiDAPH3, and we have highlighted the atoms onto which it maps.

Using the atom index and the Pipeline Pilot protocol, we can attach to each pharmacophore present in a molecule any of the CF_* descriptors mentioned above. In this case, we are talking about C-pharmacophore (CP_*) descriptors. Through this procedure, as in the fragment case, we replaced the 1 in the sequence of 0,1 bits with a real number. As you will see, to develop a model using such a huge number of bits, like those present in GpiDAPH3, was very challenging.

RESULTS

To estimate the impact of the C-fragment/pharmacophore descriptors in modeling different end points, we used them alone and in combination with 2D MOE descriptors, in techniques like Support Vector Machine (SVM)[15] and Random Forest (RF).[16] Below, we report the results obtained in modeling CYP3A4 inhibition for 1425 compounds divided in 1000 for the training set and 425 for the testing set; the experimental pIC50 (−logIC50) spans 4 log units. We used RF from the CRAN R package[17] for this purpose, especially, because it is very easy to rank the descriptors and estimate their importance.

The results reported are for the test set and they are consistent with the OOB (out of bag) results obtained for the training set. Both sets contained proprietary compounds and as a result, we have not published any structure or the associated IC50.

Table 3. Different Combinations of the CF_EState and 2D MOE Descriptors and Their Correlation with CYP3A4 IC50 for the Testing Set

	CF_EState and 2D MOE descriptors	only CF_EState descriptors	only MOE descriptors	CF_EState descriptors binary format
R^2	0.685	0.633	0.630	0.53
mean of square residuals	0.21	0.24	0.25	0.31

Table 4. Confusion Matrix for the hERG Training Set[18] Using C-Fragment and 2D MOE Descriptors

$N = 493$	predicted blockers	predicted nonblockers
true blockers (156)	135	21
true nonblockers (337)	82	255

Table 3 shows that the combination between CF_EState and 2D MOE descriptors gives a better correlation than each taken separately. In general, we have found this to be true for other data sets. From the same table, we can infer that the contribution (weight) of the 2D MOE and CF_EState descriptors, to the model that combined them, is almost equal. On the other hand, CF_EState descriptors provide better results ($R^2 = 0.633$) than their binary (fingerprint) representation ($R^2 = 0.53$). The binary representation has been obtained from CF_Estate descriptors by substituting the real numbers with the value 1, which indicates the presence of the fragment in molecule.

As can be seen from Figure 4, many of the CF_EState_* are among the top 30. As a result, we conclude that they play an important role in modeling CYP3A4 inhibition.

This was a simple way to show the potential of the C-fragment descriptors. The two horizontal axes from Figure 4 measure the variable importance. The "IncNodePurity" axe, measures the total decrease in node impurities from splitting on the variable, averaged over all trees. For classification, the node impurity is measured by the Gini index. For

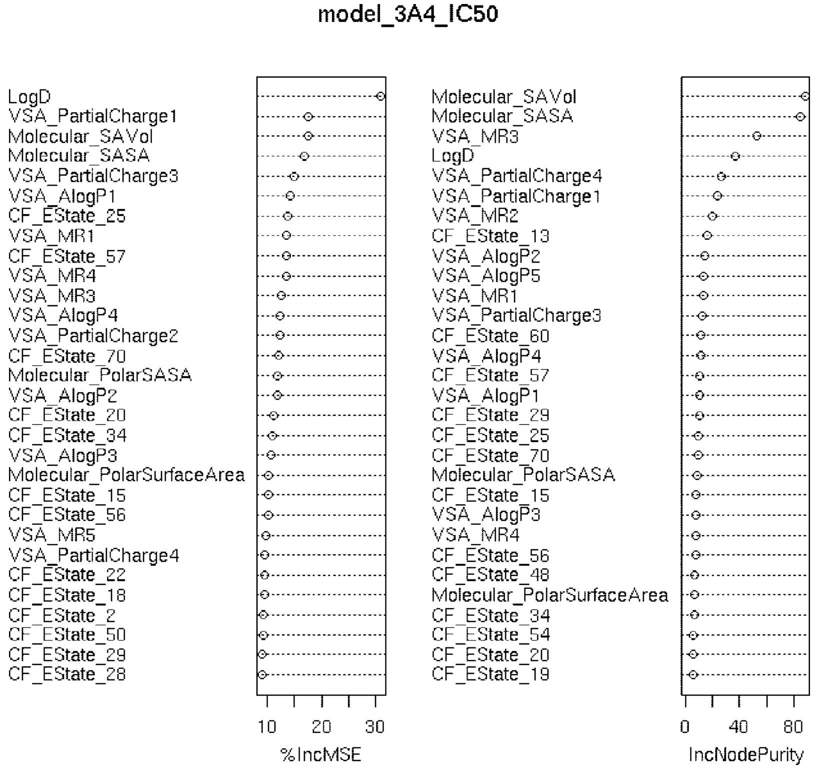


Figure 4. Rank ordering of some of the descriptors used as input for CYP3A4 model; the top descriptors are more important.

regression, it is measured by residual sum of squares. The "%IncMSE" values are calculated using the following algorithm. For each tree, the prediction accuracy on the out-of-bag portion of data is recorded. Then the same is done after permuting each predictor variable. The difference between the two accuracies are then averaged over all the trees, and normalized by the standard error.

Because recently a paper about hERG,[18] accompanied by a great deal of experimental data and structures, has come to our attention, we decided to use C-fragment/pharmacophore descriptors in modeling this end point. We used the paper's list of 561 compounds (see the Supporting Information for ref 18) made up of 495 compounds used for training and 66 compounds for testing. First, we developed a classification model using 40 μM as a threshold; we did not use the other threshold values because the purpose was only to estimate the usefulness of the newly introduced descriptors.

In 495 structures, we found 493 to be unique and we used the prevalent species at pH 7.4 ("Ionize Molecule" component from Pipeline Pilot) for C-fragment descriptors calculation. Some of the generated fragments have been condensed using symbols like A (any atom except hydrogen), Q (heavy atoms except C), single/double bond, X (halogen atoms); 47 fragments (available by request) were used in building the model with Random Forest (Random Forest from the CRAN R package[17] was implemented in Pipeline Pilot, version 6). As descriptors we used CF_Estate_*, CF_VSA_P_*, CF_VSA_N_*, CF_GA_P_*, and CF_GA_N_* in combination with some 2D MOE descriptors. The paper[18] mentioned that RF and SVM gave similar results. Therefore, any enhancement observed in our results can be attributed to the descriptors used, specifically to C-fragment/pharmacophore descriptors.

The following parameters were used in RF: mtry =15, ntree=500, weight for hERG blockers = 156 and for nonblockers=90.

The results are as follows:
OOB estimate of error rate = 20.89%
error rate for blockers 13.5% and for nonblockers 24.3%.
From the confusion matrix (see Table 4) the following values were determined:
accuracy = 79.1%
sensitivity (true blockers) = 135/156 = 86.5%
specificity (true nonblockers) = 255/337 = 75.7%
precision (true blockers) = 135/217 = 62.2%
precision (true nonblockers) = 255/276 = 92.3%
$\kappa = 0.56$

When we used only the C-fragment descriptors, the OOB error rate (of the best model) was 29.01% and $\kappa = 0.41$. For the best model developed using 2D MOE descriptors alone, the OOB error rate was 25.04% and $\kappa = 0.48$. From these results we can conclude that, in modeling hERG, the 2D MOE descriptors have a larger contribution than C-fragment descriptors.

Only the model based on the combination between C-fragment and 2D MOE descriptors was tested on the PubChem bioassay database (http://pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=376); the set contained 1702 nonblockers and 193 blockers (we removed 57 compounds defined as openers from the total number of blockers).

The model predicted 105 blockers from 193 and 1408 nonblockers from 1702 (see Table 5). These results are comparable with those reported in the literature:[18] 107 of 187 blockers were predicted correctly, and 1271 of 1690 nonblockers were predicted correctly.

Although $\kappa = 0.25$ calculated from the above confusion matrix, is slightly better than that calculated on the literature

Table 5. Confusion Matrix for the hERG Testing Set[18] using C-Fragment and 2D MOE Descriptors

$N = 1895$	predicted blockers	predicted nonblockers
true blockers (193)	105	88
true nonblockers (1702)	294	1408

Table 6. Confusion Matrix for the hERG Testing Set[18] using C-Pharmacophore GpiDAPH3

$N = 1895$	predicted blockers	predicted nonblockers
true blockers (193)	78	115
true nonblockers (1702)	295	1407

data, $\kappa = 0.18$, there is a large gap between the κ on training (0.56) and testing (0.25) sets. One possible explanation for this could be the fact that the number of fragments generated from the training set depends on its size and composition. We cannot, therefore, predict how well these fragments extend over the test set. One way to overcome this inconvenience is to use a fixed dictionary, for example, MDL public keys. In Pipeline Pilot, there is an implementation of those 166 keys known as MDLQueries.[14] Sometimes, it is useful to combine a fixed dictionary with the fragments generated using a certain scheme, but we have not applied this here because it was beyond the scope of this paper.

In addition to the above classification model, we also developed a continuous hERG model on the basis of published IC50 data. We selected all the compounds with IC50 values reported in ref 18 and removed those from the Wombat database; we ended up with 195 compounds that range from 1.25 to 9.0 log units. Using RF (regression component) with mtry = 80, ntree = 500, we obtained $r^2 = 0.38$ compared to $r^2 = 0.34$ reported in ref 18 (for 192 compounds). We have not tried to improve the model, but when we added the 66 Wombat compounds (experimental data seams to come from the same source) to the initial set the correlation coefficient increased to $r^2 = 0.43$. Further, when the model was developed only on the 66 Wombat compounds, the correlation coefficient improved to $r^2 = 0.66$.

This finding is in agreement with our observation (on in-house compounds) that the hERG model(s) are very sensitive to the type of assay (e.g., patch versus others) and that we cannot mix the same assay data if it comes from different sources.

We have observed that the fragment descriptors have a better potential (in the sense that their presence enhances the model) in the continuous (regression) models over those of classification, but more studies are necessary before generalizing this conclusion. We attribute this observation to the type of data used; in the continuous models, we used IC50 values and in the classification models, %Inhibition. Like as seen in the mathematical treatment of standard deviation, using numbers with smaller standard deviation results in answers with smaller error. Here, descriptors with higher discrimination power can impact more, well-defined and well-discriminate data like IC50.

In modeling different ADME/toxicity end points, we also used normalized C-fragment descriptors (CF_* divided by the number of heavy atoms from their composition), but we have not obtained any improvement in correlation. The fragmentation scheme, implicitly the fragments size and their frequency (distribution) are factors that could play an important role in deciding whether to use the normalized C-fragment descriptors; further investigation is needed.

C-fragment descriptors, to a certain extent, allow us to interpret the model better especially if they are ranked in the top ones. Looking at their values, chemists can modify the environment of the important fragments (their atoms could be highlighted to ease the work), in such a way as to increase for example, the potency of their compounds.

In using pharmacophores to model hERG, first we needed to map the pharmacophore atoms to the atom index. To tackle this aspect, the MOE developers provided us with a couple of Scientific Vector Language (SVL) scripts based on reverse fingerprints functions,[19] modified to our needs.

The particular case of the GpiDAPH3 pharmacophore has been even more complicated because of the huge number of descriptors (fingerprints). RF is not able to handle such a matrix (493 × 262144) and among SVM techniques, only SVM^{perf} is fast enough for large training sets.[20] Because of size of the fingerprint, we have attached to each molecule only one type of descriptor and our choice was CF_AlogP_* (logP plays an important role in modeling hERG). Another alternative is to attach a descriptor that reflects atom's role, for example, logP for hydrophobic atoms, charge for polar atoms, etc., to each atom involved in pharmacophore and not to sum but in this case the number of the descriptors will triple (each pharmacophore includes 3 atoms for GpiDAPH3 case). A complication we have been confronted with is the fact that some bits have multiple instances (states); this means that a bit can be mapped in multiple pharmacophores and that their number is sometimes quite large. We associated only one pharmacophore to each bit, the first state, because it was the most convenient way, from programming point of view, to take the output from MOE as input for our Pipeline Pilot protocol. Because of this fact is it likely to lose information or to add something that is not too valuable. To our knowledge, there is no machine learning software or theory that has addressed this problem. The sparse matrix of 493 compounds, obtained using GpiDAPH3 pharmacophores (once again we attached only the atomic AlogP value to the mapped atoms and summed over the atoms involved in the pharmacophore), was modeled using SVM^{perf} ($c = 0.5$ and the default for the rest of the options). The model was tested on the previously mentioned test set of 1895 compounds (see Table 6).

The model works well with nonblockers but cannot be used for blockers. As a result, a predicted nonblocker has a high probability of being nonblocker. Conversely, prudence would dictate experimentally testing those compounds for which the treatment predicts are blockers. In addition to multiple instances problem, the fact that we have not used 2D MOE descriptors could explain a less performing model ($\kappa = 0.16$). On the other hand, the κ value is close enough to that calculated from literature data (κ 0.18) to foresee opportunities for C-pharmacophore descriptors in the 2D case. Finally, we should not neglect the 3D pharmacophores (one pharmacophore or a consensus of pharmacophores), as well as a combination between C-fragment and C-pharmacophore descriptors in modeling different bioactivities and ADME/toxicity end points.

CONCLUSIONS

Using a well-defined set of fragments/pharmacophores a new methodology to calculate fragment/pharmacophore descriptors for any molecule onto which at least one fragment/pharmacophore can be mapped was presented. The procedure can be applied to any descriptor defined at the atomic level such as AlogP, MR, E-State, etc. A Pipeline Pilot protocol to calculate all C-fragment descriptors presented in this study was written. To handle the pharmacophore case, SVL scripts have been used in addition to the Pipeline Pilot protocol. At present, to calculate C-fragment/pharmacophore descriptors, we only summed over the atoms making the fragment but other mathematical functions could be applied if there is reason to do it.

The real power of the C-fragment/pharmacophore descriptors comes from the fact that different molecules have different values for the same fragment because of a different molecular environment. As a consequence, we can replace the "1" value from a fingerprint with C-fragment/pharmacophore value in numerous applications like diversity/similarity analysis, virtual screening, etc.

To assess their utility, the C-fragment/pharmacophore descriptors have been used to model different ADME/toxicity end points (e.g., CYP inhibition and hERG potential). The best results were obtained when we used them in combination with some 2D MOE descriptors. In the CYP inhibition model, the C-fragment and 2D MOE descriptors had equal weights, but for hERG data, the 2D MOE descriptors appeared to have a larger contribution. The CYP data also showed the advantage of using the C-fragment descriptors over their fingerprint representation. Through the ranking technique from RF we identified that some of the C-fragment descriptors play an important role in CYP inhibition model.

The modeling of hERG data set did show a slight enhancement over literature treatments and we attribute this, at least in part, to the power of C-fragment descriptors. While somewhat disappointing, when the model was expanded to include C-pharmacophore descriptors based on GpiDAPH3, its performances decreased. Some factors that impeded the process have been identified. One of them, related to the inability of machine learning techniques to handle multiple instances, will most likely open new research areas.

More studies are needed to assess the full potential of C-fragment/pharmacophore descriptors in chemical informatics. Some of them have been suggested in this paper and we hope this will inspire others to fully expand the field.

REFERENCES AND NOTES

(1) Bender, A.; Glen, R. C. Molecular similarity: A key techniques in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204–3218.

(2) Conne, M. M.; Venkataraghavan, R.; McLafferty, F. W. Computer-aided interpretation of mass spectra. 20. Molecular structure comparison program for the identification of maximal common substructures. *J. Am. Chem. Soc.* **1977**, *99*, 7668–7671.

(3) Free, S. M.; Wilson, J. W. A mathematical contribution to structure−activity studies. *J. Med. Chem.* **1964**, *53*, 395–399.

(4) Barnard, J. M. Substructure searching methods: Old and new. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 532–538.

(5) Takahashi, Y.; Sukekawa, M.; Sasaki, S. Automatic identification of molecular similarity using reduced-graph representation of chemical structure. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 639–643.

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

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

(8) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular similarity searching using atom environments, information-based feature selection, and a naïve bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170–178.

(9) Xiang, L.; Glen, R. C. Novel Methods for the Prediction of logP, pK_a, and logD. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 796–805.

(10) Xiang, L.; Glen, R. C.; Clark, R. D. Predicting pK_a by molecular tree structured fingerprints and PLS. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 870–879.

(11) Hall, H. L.; Mohney, B. The electrotopological state: Structure information at the atomic level for molecular graphs. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 76–82.

(12) Wildman, A. S.; Crippen, M. G. Prediction of physicochemical parameters by atomic contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868–873.

(13) *Molecular Operating Environment (MOE)*, version 2008; Chemical Computing Group Inc.: Montreal, Canada, 2008.

(14) *Pipeline Pilot*, version 6.0; Accelrys: San Diego, CA, 2008.

(15) Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Min. Knowledge Discovery* **1998**, *2*, 121–167.

(16) Breiman, L. Random forest. *Machine Learning* **2001**, *45*, 5–32.

(17) The Comprehensive R Archive Network. http://cran.r-project.org (accessed March 27, 2008).

(18) Li, Q.; Jorgensen, S. F.; Oprea, T.; Brunak, S.; Taboureau, O. hERG Classification model based on a combination of support vector machine method and GRIND descriptors. *Mol. Pharm.* **2007**, *5* (1), 117–127.

(19) Williams, C.; Schreyer, S. K. *Reverse Fingerprinting and Mutual Information- Based Activity Labeling and Scoring (MIBALS)*; Chemical Computing Group Inc.: Montreal, Canada (to be published).

(20) Joachims, T. A Support Vector Method for Multivariate Performance Measures. Presented at the International Conference on Machine Learning [Online], Bonn, Germany, 2005. Support Vector Machine for Multivariate Performance Measures Web Site. http://svmlight.joachims.org (accessed Oct 13, 2008).

CI800339P