Novel 2D Fingerprints for Ligand-Based Virtual Screening

Todd Ewing,* J. Christian Baber, and Miklos Feher†

Neurocrine Biosciences, 12790 El Camino Real, San Diego, California 92130

Received April 28, 2006

This paper describes the development of a set of new 2D fingerprints for the purposes of virtual screening in a pharmaceutical environment. The new fingerprints are based on established ones: the changes in their design included the introduction of overlapping pharmacophore feature types, feature counts for pharmacophore and structural fingerprints, as well as changes in the resolution in property description for property fingerprints. The effects of each of these changes on virtual screening performance were monitored using two types of training sets, emulating different stages in the drug discovery process. The results demonstrate that these changes all lead to an improvement in virtual screening performance.

INTRODUCTION

Molecular fingerprints have become fundamental tools in ligand-based virtual screening.[1−3] Their popularity is mainly a result of their simplicity, as well as their ease of application and speed of calculation, which are critical when virtual libraries containing millions of molecules are being screened. Fingerprints represent gross simplifications of the molecule that allow rapid comparisons for the purposes of finding "similar" ones. They can be categorized according to the way this simplification is achieved as structural,[4−7] pharmacophore,[8,9] and property [10] fingerprints, although some combine these categories.[11] Fingerprints can represent 2D or 3D information; this work is exclusively concerned with the former. Although 3D methods have tremendous potential for accuracy and detail, the ambiguity of the active conformation of flexible molecules, combined with the surprisingly good results of 2D methods in previous work[12,13] led to our continued use and development of these 2D methods. The available 2D fingerprints describe molecules either by indicating the presence and absence of certain features (e.g, as bit strings of ones and zeros or as bit position vectors indicating the indices of bits which are in the "on" state) or by displaying the number of occurrences of those features in the molecule (such as through a histogram vector).

The aim of this work was to look at the effect of changes in the structure of different fingerprint types on virtual screening performance in situations that attempt to approximate lead-finding and lead-optimization scenarios. In particular, three effects were studied. First, our hypothesis was tested as to whether the use of overlapping pharmacophoric atom types would improve virtual screening performance (in this scheme, molecular features may belong to multiple categories). Second, the effect of replacing fingerprint bits (absence or presence) with counts (number of occurrences) on virtual screening performance was studied. Third, the effect of increasing the resolution in property fingerprints was tested. It was hoped that these changes in fingerprint structure would translate to better virtual screening performance.

METHODS

All molecule manipulations were performed using the MOE suite of programs.[14] Molecules were first protonated/deprotonated to ensure consistent representation but otherwise kept in a 2D representation. The process for selecting training and validation sets has been described in detail.[13] Briefly, antagonists for four different G-protein-coupled receptors (GPCRs) were considered: corticotropin-releasing factor receptor-1 (CRF), gonadotropin releasing hormone receptor (GnRH), melanocortin receptor-4 (MC4), and melanin-concentrating hormone receptor (MCH). Two different kinds of training sets were applied for these receptors: "ordinary" and "hard" sets. In the ordinary sets, the actual number of compounds was chosen on the basis of the number available both in-house and from the literature. The target number of molecules was 1000 for the training sets and 2000 for the validation sets. Molecules were selected into these training sets or designated as "active" in the validation sets with K_i values of less than 100 nM. In contrast, the hard training sets contained only 20 molecules, and the corresponding validation sets contained actives with K_i values of less than 1 μM. Molecules with less than 50% inhibition at a 10 μM concentration in binding assays were defined as inactives. All inactive structures in the validation sets were assembled using molecules generated at Neurocrine within the project and targeting the given receptor. The selection of training and test sets was based on structural diversity and involved the MACCS fingerprint. This method of set selection is the same as that described in our previous work.[13] The use of known inactives selected in this manner should be a more challenging test than the use of decoy sets that are known to have a number of potential issues such as the assumption of inactivity for compounds that may in fact bind at the given target and artificially high enrichment resulting from differences in 1D properties[15] or the presence of grossly different chemotypes.[16]

* Corresponding author phone: (858) 617-7299; fax: (858) 617-7619; e-mail: tewing@neurocrine.com.

† Senior author. Current address: The Campbell Family Institute for Breast Cancer Research, University Health Network, Toronto Medical Discovery Tower, 101 College Street, Suite 5-361, Toronto, Ontario M5G 1L7, Canada; phone: (416) 581-7611; e-mail: mfeher@uhnres.utoronto.ca.

10.1021/ci060155b CCC: $33.50
Published on Web 09/16/2006

Virtual screening was performed by scoring each validation set molecule using the given fingerprint and on the basis of its Tanimoto similarity to its nearest neighbor in the training set. In such virtual screening runs, similarities between the molecule in question and all molecules of the training set are calculated, with the highest score accepted as the final score of the compound (known as the MAX fusion rule).[17] The performance of different fingerprints were compared using plots of the true positive rate as a function of the false positive rate for different score thresholds, using the so-called ROC (receiver operating characteristic) plots. Instead of using enrichments at given recovery rates, we describe the overall performance of the applied methods using the following quantity

$$\text{performance} = 100 \times \frac{\text{area under the ROC curve and over the diagonal}}{\text{area over the diagonal}} \qquad (1)$$

The overall performance defined in this manner is 100% for an ideal behavior (all actives recovered before any of the inactives), 0% for a completely random behavior, and negative for worse than random behavior (i.e., it is normalized between ±100%). The overall performance provides a measure of the capability of a fingerprint without having the potential bias introduced when a specific level of recovery is selected. In practice, when the sampling is 10% or less, a 3% increase in the recovery of actives coincides with about a 1% increase in the overall performance, based on the examples shown in Table 3b, in which active recovery improvements from 10 to 30% contributed to overall performance improvements from 3 to 10%. As the proportion of the set sampled increases above 10%, the recovery of actives tends to saturate, and the relative differences tend to diminish.

FINGERPRINT DEVELOPMENT

All the fingerprints described in this work were developed using the SVL language within the MOE suite of programs.[14] The following fingerprints, available within MOE, were investigated and modified in this work: MACCS,[4,14] TGT,[14] and MP61.[10] Briefly, the MACCS fingerprint (also called 166-bit MDL keys) was originally developed for database searching.[4] Its bits represent the presence in the molecule of certain atom types, bond types, atom environments, groups, and properties. The typed graph triangle (TGT) fingerprint is a three-point pharmacophore fingerprint calculated from a 2D molecular graph.[14] Each atom is represented using one of the following type definitions: donor (which also includes cations), acceptor (which also includes anions), polar (which includes atoms that are simultaneous acceptors and donors, such as a hydroxyl group), and other (hydrophobes and unassigned features such as polar heteroatoms that do not participate in hydrogen bonding). All possible triplets from these typed atoms, using these pharmacophore types, are coded using graph distances, and the triangles of no interest (e.g., those containing three hydrophobes) are discarded. The rest are sorted and binned with the resulting fingerprint being represented as a sparse feature list. The MP61 fingerprint[10] is a molecular property-based fingerprint using 61 binary encoded descriptors. The encoded properties include atom counts, bond counts, topological, van der Waals surface and charge descriptors, log *P*, and complex surface area descriptors; these were carefully selected to eliminate correlating ones. For each of these properties, the bits are turned on if the value for a given molecule is above the median value established for that property using a database of ~1.3 million compounds from medicinal chemistry vendors. We have recently shown[13] that all these fingerprints perform well in practical virtual screening. Next, the fingerprints developed in this present work will be described.

Table 1. TGTO Representation of Overlapping Types and the Similarities between Them[a]

types {overlaps}	cation {+, D}	donor {D}	polar {D, A}	acceptor {A}	anion {−, A}	hydrophobe {H}	untyped {X}
cation {+, D}	1	0.5	0.33				
donor {D}	0.5	1	0.5				
polar {D, A}	0.33	0.5	1	0.5	0.33		
acceptor {A}			0.5	1	0.5		
anion {−, A}			0.33	0.5	1		
hydrophobe {H}						1	
untyped {X}							0

[a] The top row and first column show pharmacophore types and how each is represented with overlapping categories. The intersection of rows and columns show the average Tanimoto similarity resulting from matching triplets that differ only by the given pharmacophore types. Blank cells represent zero values.

Pharmacophore Fingerprint with Overlapping Atom Types (TGTO). As with TGT, typed graph triangles were built using different atom types and standard Tanimoto similarity to evaluate the similarity of molecules on the basis of pharmacophore triplets. However, to improve the performance of the TGT fingerprint, the following changes were made. First, the number of feature types was increased to include the following categories: cation, donor, anion, acceptor, hydrophobic, and other (unassigned). The novelty of the definition is that an atom may belong to more than one category (e.g., acids are both anions and acceptors, polar atoms are both acceptors and donors, etc. (see Table 1)). This is in contrast to usual type definitions, such as those in TGT, that are mutually exclusive. For example, if an atom is typed as polar (simultaneous acceptor and donor) in the TGT fingerprint, it has no similarity to either acceptors or donors, whereas in the TGTO fingerprint these definitions overlap (e.g., an acceptor has a 50% overlap with a polar atom). Overlap was implemented for polar pharmacophore types by replacing each triplet containing a polar with two equivalent triplets, one representing the polar as an acceptor and one representing the polar as a donor. Similarly, overlap for anions (and cations) was implemented by duplicating each triplet containing an anion (or cation) with an equivalent triplet representing the anion (or cation) as an acceptor (or donor). Second, the definition of hydrophobes was changed, excluding those atoms adjacent to atoms already typed as anion, cation, donor, acceptor, or polar and also atoms that are fully buried (similar to the definition of "grease" in MOE). Third, unlike TGT, which lumps hydrophobic atoms together with those of type other, TGTO only considers hydrophobes and excludes atoms of type other from the analysis.

Pharmacophore Fingerprint with Feature Counts (TGTF). For this fingerprint, a different kind of modification was undertaken. In the standard TGT fingerprint, the bits represent the presence or absence of typed triangles with given feature combinations and topological interfeature distances. In TGTF, the fingerprint was modified to record the number of times each feature occurs in a molecule. The fingerprint elements and the counts were stored together as a single vector of integers, the first half of the vector corresponding to the hash index of each feature and the second half corresponding to the count of each feature. A similarity metric analogous to the Tanimoto similarity and originally developed for histograms[18] was defined for molecule comparisons, on the basis of the number of features shared by two molecules, c_{ij}, normalized by the average number of features present in both molecules, n_i and n_j, as follows

$$t_{ij} = \frac{2c_{ij}}{n_i + n_j} \tag{2}$$

An additional similarity measure based on the mean similarity between features occupied in either molecule was also tested. However, in initial testing, this was found to perform worse than the standard measure and thus was not used any further. Fingerprints involving feature counts are rare but not unprecedented;[19] however, a study of the effect of such a change on virtual screening has not been described to our knowledge.

Pharmacophore Fingerprint with Overlapping Types and Feature Counts (TGTFO). The two sets of modifications, described above for defining the TGTO and TGTF fingerprints from the standard TGT, were applied simultaneously in the TGTFO fingerprint. Because this fingerprint contains actual feature counts, the histogram Tanimoto similarity as defined above was used for scoring similarity.

MACCS Fingerprint with Feature Counts (MACCSF). In addition to the bit elements showing the availability of certain structural features, the number of occurrences for such features is stored in the fingerprint. Molecular similarity is determined from these fingerprints using the histogram Tanimoto metric (see above).

Property Fingerprint Using Range Quartiles (MP61q). As mentioned above, the MP61 fingerprint bits are determined on the basis of whether the property value is above or below the median value.[10] When initiating this work, we saw two potential opportunities to improve the virtual screening performance of this fingerprint. First, instead of subdividing the property range into two classes (above or below the median), it was decided that the range should be divided into four quartiles. Second, instead of using vendor catalogs to define the property ranges for the fingerprint,[10] it was thought that the MDDR database,[20] containing a collection of nearly 150 000 pharmaceutically relevant compounds, would be more appropriate. Although the number of compounds in this collection is significantly smaller than that used in defining MP61, it was hoped that the quality and druglikeness of the compounds in this collection would make up for this difference given our specific application. Also, because of the wide availability of the MDDR collection, it is expected that any updates or further developments would be easier with this collection.

To use the original MP61 fingerprint in virtual screening, yet another extension of the standard Tanimoto similarity, average Tanimoto metric, was applied.[10] This was necessary because in the MP61 fingerprint both the 1's and the 0's are equally important, showing whether the molecule for a given property is above or below the median, and the important value in this case is the average of the Tanimoto coefficients calculated for the two bits.[10] (It must be noted that the code as originally published[21] had an erroneous implementation of the averaged Tanimoto similarity, which was corrected in the present work.) The average Tanimoto metric was also applied for the MP61q fingerprint and the bit definitions in MP61q were defined to accommodate such calculations (see Table 2). The original MP61 fingerprint uses a single bit to indicate whether the value of a given property is above or below the median value for that property. For the expansion to quartiles, it was necessary to use three bits to indicate in which quartile the current value belongs. However, to allow the average Tanimoto coefficient to measure similarity, bits were set on the basis of whether the value being considered was greater than the value determined for the first, second, and third quartiles, respectively, rather than only indicating presence in a given quartile. Thus, for a value in the third quartile, the first two bits would be set, indicating that the value is greater than the top of the first and second quartiles, respectively, and the third bit would remain unset. Although this significantly increases the length of the fingerprint, the new fingerprint is still 183 bits long with the 61 properties considered, making is substantially shorter than, for example, the pharmacophore fingerprints applied in this work.

Table 2. MP61q Representation of Quartiles and Similarities between Them[a]

quartile	1st	2nd	3rd	4th
bit representation	000	100	110	111
1st, 000	1	0.67	0.33	0
2nd, 100	0.67	1	0.67	0.33
3rd, 110	0.33	0.67	1	0.67
4th, 111	0	0.33	0.67	1

[a] The top row and first column show the bits set for values in a given quartile. The intersection of rows and columns show the average Tanimoto similarity assigned to a given pair of categorized values.

RESULTS AND DISCUSSION

The performance comparison for the studied fingerprints is presented in Table 3. Table 3a provides values for the ordinary case, which was put together using data from four medicinal chemistry projects at Neurocrine. Each project had a large amount of data available, and the training and test set selection proceeded using the procedures described above. Because of the careful selection process (it was ensured that each chemical class represented in the test set was also available in the training set) and the fact that the active category contained molecules with better than 100 nM activity, the observed enrichments were high and overall performance values were close to the theoretical values. We believe that this approximates a lead optimization scenario, when wealth of data is available to drive the computational predictions. Table 3b shows the performance for the hard case, in which the training set was small, the test set might contain structural classes that are not fully covered in the training set, and there is a lower threshold for the selection of actives in the test set (1 μM) than in the previous case. This, in fact, somewhat approximates new lead finding in

Table 3. Overall Percentage Performance of a Combination of Different Virtual Screening Approaches on the Validation Sets[a]

(a) ordinary case

	CRF	MCH	MC4	GNRH	mean	range
MACCS	98.8	98.0	100.0	99.2	99.0	2.0
TGT	98.1	97.8	99.2	98.5	98.4	1.4
MP61	94.4	96.3	99.3	96.6	96.7	4.9
MACCSF	98.8	98.0	100.0	99.6	99.1	2.0
TGTO	98.3	97.9	99.6	99.0	98.7	1.7
TGTF	98.5	98.2	99.6	99.6	99.0	1.4
TGTFO	98.7	98.9	99.7	99.6	99.2	1.0
MP61q	98.1	97.1	99.5	97.7	98.1	2.4

(b) hard case

	CRF	MCH	MC4	GNRH	mean	range
MACCS	94.1	72.6	94.8	66.9	82.1	27.9
TGT	90.5	66.4	88.0	52.4	74.3	38.1
MP61	65.5	61.7	84.1	53.5	66.2	30.6
MACCSF	97.0	79.8	96.1	66.4	84.8	30.6
TGTO	91.2	62.5	93.6	64.1	77.9	31.1
TGTF	93.4	72.7	92.0	62.2	80.1	31.2
TGTFO	94.1	76.5	95.8	73.9	85.1	21.9
MP61q	82.0	57.6	90.0	56.2	71.5	33.8
MP61(mddr)	65.2	65.6	89.3	48.9	67.2	40.4

[a] Overall performance, expressed as a percentage of the area under the ROC curve but above the diagonal, relative to the total area above the diagonal, as defined by eq 1. For the description of the symbols representing different methods and all other information, see text.

drug discovery (especially in search for a follow-up compound or in a lead-hopping scenario to avoid PK or IP issues). In this case, the virtual library that is screened may indeed contain novel structural classes, and the active molecules are fewer in number and have lower activities. Not surprisingly, the performance of all methods is substantially inferior for the hard case. As described previously,[13] out of the three established fingerprints, MACCS performs consistently best, followed by the pharmacophore fingerprint TGT and the property fingerprint MP61. However, it has been shown that fingerprints based primarily on structural features, such as MACCS, are generally less useful when hits structurally different from the training set are sought (lead hopping or scaffold hopping), and in such cases, pharmacophore and property fingerprints are more useful.[13,22] It is for this reason, combined with the fact that there appears to be more room for improvement, that effort has been concentrated on modifying the TGT and MP61 fingerprints rather than making major changes to the MACCS fingerprint.

The performance is also shown in Figures 1 and 2. Figure 1 displays a classical ROC plot with the performance averaged over the four receptor systems studied. Ideal performance in this plot would involve recovering all actives before the first inactive is found, which would correspond to a vertical line to 0.1% and then a horizontal line at 100%; the closer the curve is to such a line, the better the performance. Figure 2, on the other hand, presents performance improvement as a function of the percentage of actives recovered, whereby ideal performance would be represented by the diagonal. In different drug discovery scenarios, the objectives might be different: in some cases, the only important thing is to turn up new actives (which could correspond to, for example, 40−60% recovery); in other cases, as many actives as possible need to be found (i.e., a high recovery rate of >80% might be sought). As can be seen in Figure 2, some of the curves intersect each other (i.e., the relative performance of different methods will be different at different recovery rates). In contrast, the numbers presented in Table 3 relate to the area under the ROC curve (i.e., provide overall performance across all recovery rates). Next, we will look at the performance of the new fingerprints.

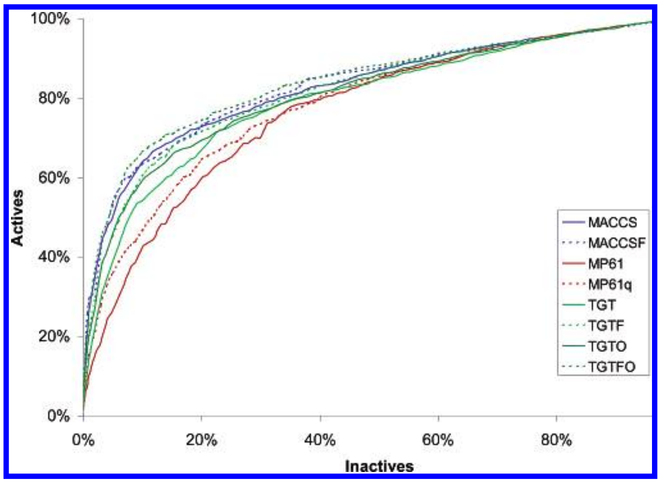

Figure 1. ROC plot for the hard test set for each of the fingerprints averaged across the four targets.

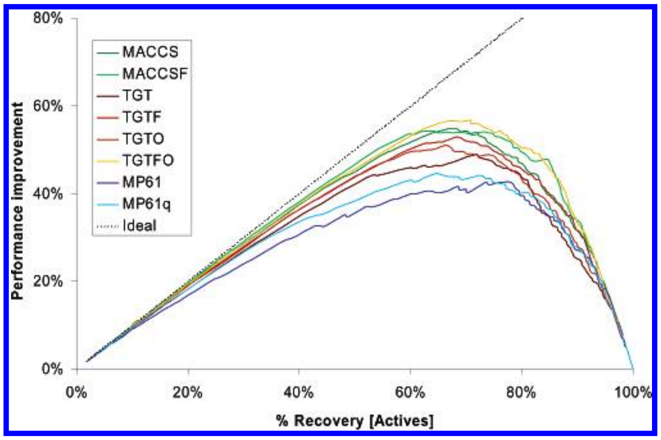

Figure 2. Performance improvement over random performance as a function of the recovery rate of actives for different fingerprints. The performance improvement is measured as the difference between the percentage recovery of actives and inactives.

Pharmacophore Fingerprints. The new pharmacophore fingerprints performed better than the original typed graph fingerprint (TGT) with a single exception. In almost all cases, the introduction of overlapping types and feature counts led to independent performance improvements, and when both were applied (as in TGTFO), the improvements were almost additive. In terms of the average overall performance, the improvements upon introduction of the overlapping types, feature counts, or both for the ordinary set are 0.3, 0.6, and 0.8%, respectively, whereas for the hard case, the corresponding values were 3.5, 5.8, and 10.8%. Clearly, the changes in these fingerprints appear to have a greater effect for the hard scenario than for the ordinary one; of course, the room for improvement is also much greater in the hard case.

Property Fingerprints. The division of the property range into four quartiles (as in MP61q) rather than just two halves (as in MP61) seems to lead to a consistent improvement,

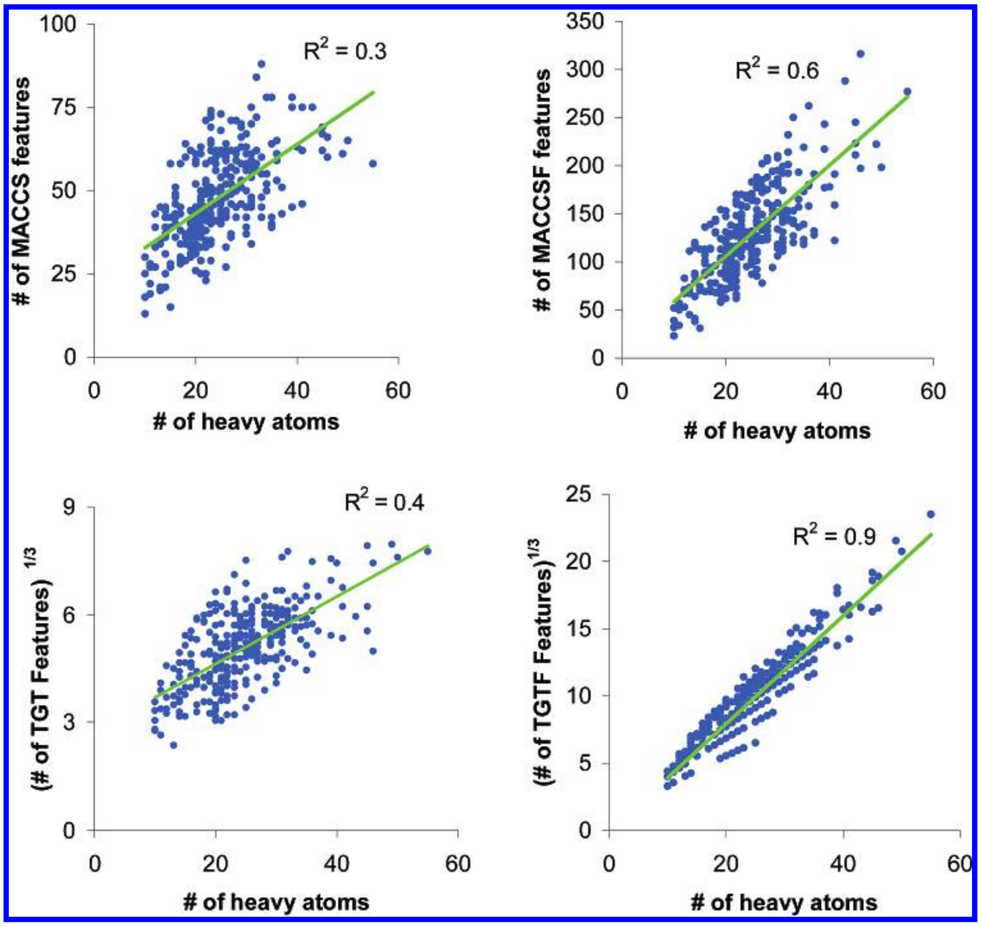

Figure 3. Relationship of the number of fingerprint features with molecular size. Each point represents one molecule from a set of 306 orally active drugs, selected to represent a broad range of drug molecule size and properties. The number of fingerprint features for each molecule is plotted against the number of heavy atoms. For the TGT and TGTF fingerprints, the cube root dependence is plotted, because the number of pharmacophore triplets is proportional to the number of atoms, raised to the third power.

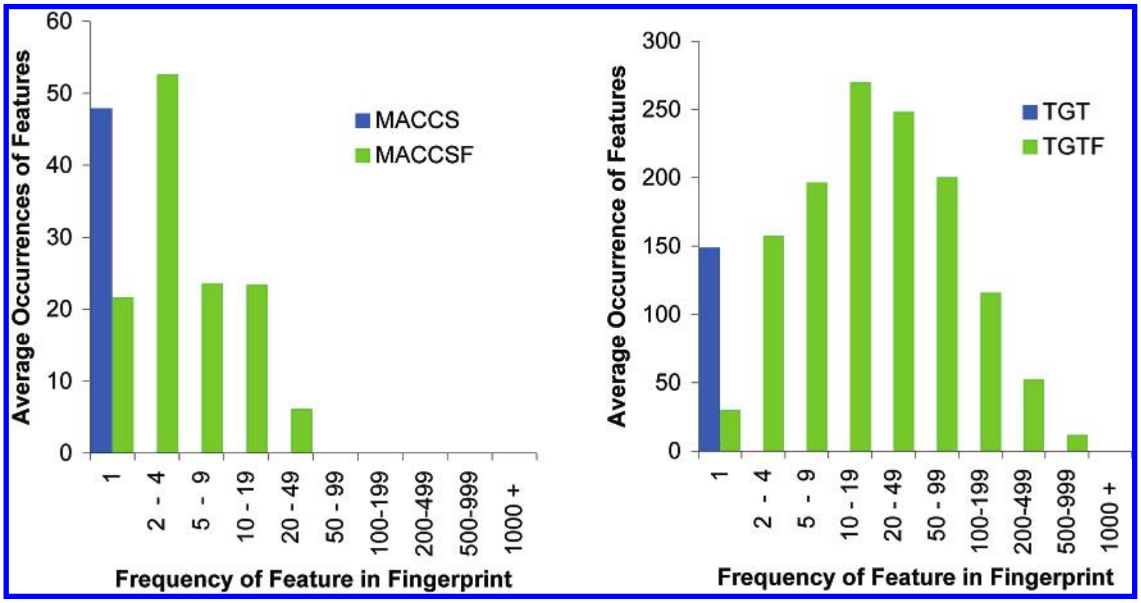

Figure 4. Distribution of feature frequencies for different fingerprints, as determined from a set of 306 orally active drugs. The features were binned according to the frequency, or the total number of repeat occurrences, of the feature in a molecule. Since the MACCS and TGT fingerprints discard repeat occurrences of features, the frequencies are always one. The bin totals were averaged over all the molecules, so that the histograms above represent a typical distributions for a druglike molecule. Summing over all bins gives the typical number of feature occurrences per molecule, which are MACCS 48, MACCSF 130, TGT 150, and TGTF 1300. The increased width of the TGTF distribution is partly caused by the cubic dependence of the number of atom triplets on molecule size.

with the only slight decrease in performance occurring in the hard case for the MCH receptor. The average improvement when property quartiles are used is 0.1 and 2.7% for the ordinary and hard cases, respectively. This kind of improvement was also observed by us for other receptors not discussed here.

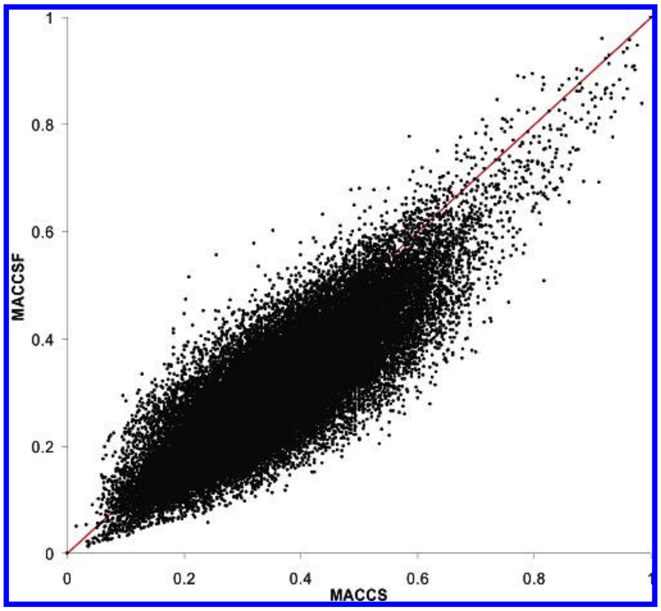

Figure 5. Correlation plot for the Tanimoto similarities, generated for all possible molecule pairs of compounds using 306 orally available drugs and the MACCS and MACCSF fingerprints.

Table 4. Statistics for Related Fingerprint Pairs[a]

	MACCS−MACCSF	TGT−TGTF	TGT−TGTO	TGT−TGTFO	MP61−MP61q
r^2	0.70	0.81	0.41	0.33	0.69
std error	0.07	0.08	0.13	0.14	0.05
mean (1)	0.37	0.60	0.60	0.60	0.65
mean (2)	0.31	0.50	0.53	0.45	0.65
Δmean	0.06	0.09	0.06	0.15	0.00
median (1)	0.36	0.62	0.62	0.62	0.64
median (2)	0.30	0.53	0.55	0.46	0.65
95 percentile (1)	0.58	0.80	0.80	0.80	0.79
95 percentile (2)	0.50	0.73	0.72	0.65	0.78

[a] As determined for pairs of fingerprints on 306 orally available drugs by calculation of Tanimoto similarities for all possible molecule pairs. Numbers in parentheses denote whether the property relates to the first or second fingerprint in the first row of the table.

As described above, the MP61 and MP61q fingerprints differ in both the resolution in the description of properties, as well as the set used to define such property ranges. To separate the two effects, we redefined the original MP61 fingerprint using the MDDR reference set. The results for the hard case are displayed in Table 3b under the name MP61(mddr). It appears from the data that the use of the MDDR reference set led to modest average improvements in performance (∼1%) but with large variations. Although, on the basis of the averages, we can conclude that the general performance improvement from MP61 to MP61q is caused by the higher property resolution to a greater extent and less to the set used to define the ranges, this conclusion does not necessarily hold for the individual datasets that might even show opposite trends. Despite this, we can conclude that the introduction of MP61q instead of MP61 generally improves performance, and the use of the MDDR set to define the property ranges at least makes the work more easily reproducible.

It is difficult to compare our results to those in the literature.[11] Xue et al. compared two mini-fingerprints: one that included both a structural and a property descriptor with 8 property ranges (SE-MFP) and another with structural and binary property descriptors (MP-MFP), with the property part of the latter fingerprint being identical to MP61. In the validation tests on a set of 549 active molecules belonging to 38 different activity classes and 5000 decoys, the MP-MFP fingerprint performed slightly better than SE-MFP. Because the structural part of the fingerprints was identical, this improvement can be attributed to the property part. Unfortunately, the property parts differed not only in the way the properties were represented by ranges versus medians but also in the number of descriptors applied (14 properties in SE-MFP vs 61 properties in MP-MFP). Thus, on the basis of the reported results, a direct comparison of the two fingerprints is not feasible. From the current work, it appears likely that, if the number of properties had been increased in ref 11 for the SE-MFP fingerprint but the range description had been retained, the performance might have been better than that observed for MP-MFP.

Structural Fingerprints. The MACCS fingerprints are widely used and have been found to be among the best 2D fingerprints, even surpassing 3D search methods.[12,23,24] The public version of the MACCS fingerprints, applied in this work, has not been optimized for virtual screening. It appears from the current results that the addition of feature counts to the MACCS fingerprints leads to significant improvements for some receptors but has little effect for others. The average improvements of 0.1 and 2.7% for the ordinary and hard test sets, respectively, cover four cases with little or no improvement. In summary, it appears that using counts for different structural features (as in MACCSF) instead of the presence and absence bits (as in the MACCS fingerprint) is generally either beneficial or neutral, so it is likely to be preferable in virtual screening to the traditional MACCS fingerprints.

Size Effects in Frequency Fingerprints. The numbers in Table 3 indicate the gross performance of the pharmacophore fingerprints but hide a major potential difference. The fingerprints indicating only the existence of certain groups of typed triangles (such as TGT and TGTO) are expected to be less affected by the number of mappings of a given pharmacophore triplet than the feature-counting pharmacophores (such as TGTF and TGTFO). A similar size-dependent difference in performance is also expected between the structural fingerprints MACCS and MACCSF. This could appear because frequency fingerprints encode the size of the molecule directly, whereas binary fingerprints only encode the size indirectly through the fact that larger molecules tend to have a greater variety of features. In large molecules many kinds of typed distances invariably occur and hence the fingerprint indicating the mere presence and absence of features should lead to worse performance than if the actual number of such features is considered.

To test this hypothesis, further studies were undertaken on a representative sample of 306 orally active drug molecules. Since the definition of the Tanimoto similarity includes the number of shared and unique features, if a fingerprint were to encode molecular size, the total number of features encoded for a molecule should correlate with the size of the molecule to a first level of approximation. The results, displayed in Figure 3 for different fingerprints, indeed indicate some size-dependence. Clearly, the size of the MACCS and TGT fingerprints only weakly correlate with molecular size, whereas the MACCSF and TGTF fingerprints

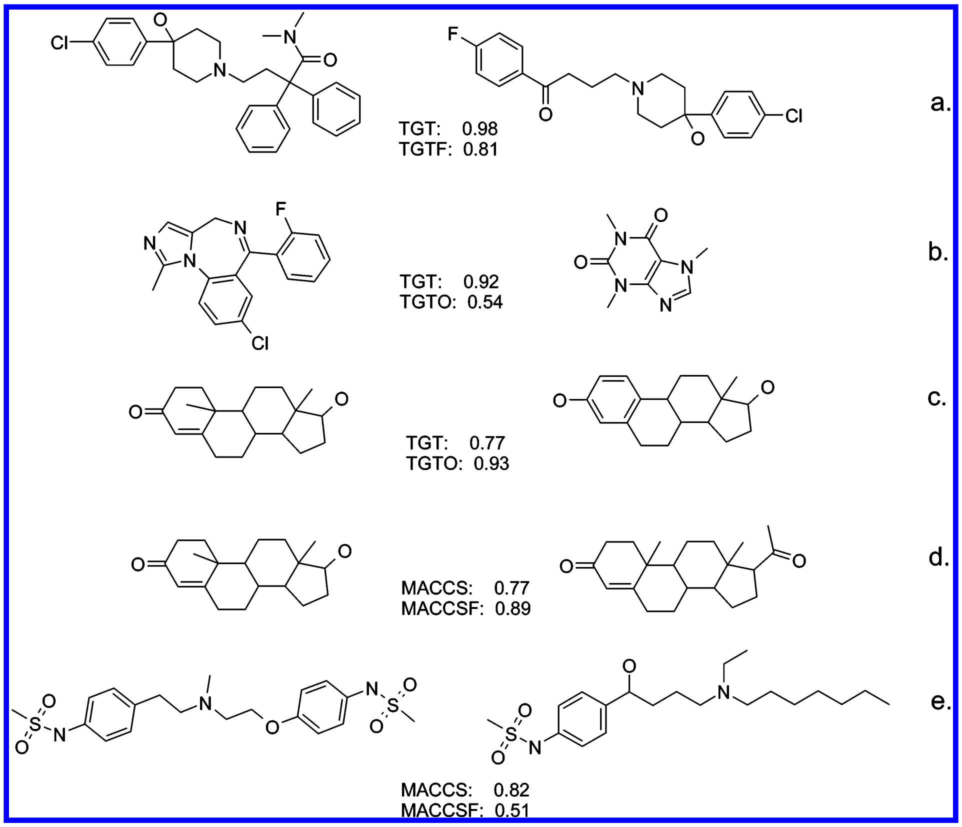

Figure 6. Illustrative molecule pairings highlighting differences between related fingerprints. (a) Loperamide and haloperidol have high TGT similarity but the two molecules appear less similar when the number of occurrences is considered for pharmacophore types. (b) The high TGT similarity of midazolam and caffeine is caused by the number of unassigned atom types that behave as hydrophobes in TGT. (c) The testosterone−estradiole pair appears to be highly nonsimilar when using TGT because the pharmacophore definitions for the carbonyl and hydroxyl groups do not overlap. (d) Because of the number of type repetitions that are neglected, the MACCS fingerprint in the testosterone−progesterone pair causes the hydroxyl−methyl-keto change to be accentuated leading to low similarity. (e) The repetition of the phenyl sulfonamide moiety remains unaccounted for in the dofetilide−ibutilide pair, causing higher similarity in MACCS than should be warranted.

correlate more strongly with the number of heavy atoms. In addition, the *y* intercepts for the MACCS and TGT correlations deviate much farther from the origin than the MACCSF and TGTF fingerprint correlations, further indicating a rather flat relationship with size for these molecules. This is understandable, given that the MACCS and TGT fingerprints were constructed to eliminate redundant feature information to compress the size of the fingerprints and increase the speed of similarity comparisons, with the assumption that the redundant features added little to the molecular description. (In a similar manner, over-abundant features also add little to molecular description.) In our implementation of MACCSF and TGTF, the fingerprint size and speed costs approximately double, since a second vector of information is needed to store and compare feature counts. It is clear from this analysis that to a first level of approximation, retaining a record of the repeated fingerprint features should improve the ability of the fingerprint to capture the size differences between molecules.

Frequency Contributions of Fingerprint Features. The frequency count of a fingerprint feature indicates the importance of that feature's contribution to the similarity comparison. Correspondingly, the range of frequency counts indicates how differently the frequency fingerprint will behave compared to a standard fingerprint (i.e., if the frequency counts are uniform across all features, then the frequency-weighted fingerprint will be indistinguishable from the standard fingerprint). To assess the range of frequencies encoded for an average druglike molecule, we further analyzed the fingerprints for the set of 306 orally active drugs, as shown in Figure 4. The MACCS and TGT fingerprints are composed entirely of singletons by the nature of their construction. The MACCSF fingerprint typically contains a small fraction of singletons with the majority of features repeated between two to nine times and a small fraction of features even being repeated more than 20 times. As a result, when redundant features are discarded to form the MACCS fingerprints, more than half of the approximately 130 MACCSF features of an average oral drug molecule are eliminated to form the approximately 48 unique MACCS features of a typical molecule. The TGTF fingerprint typically contains a very small fraction of singletons with the majority of features repeated between 5 and 99 times. Consequently, when redundant features are discarded to form the TGT fingerprint, the vast bulk of the approximately 1300 TGTF features, on average, are discarded to arrive at the approximately 150 unique TGT features of an average molecule.

Comparison of Results with Different Fingerprints. The performance improvement in a virtual screening exercise on using the fingerprints developed in this work has been illustrated above. However, when different fingerprint results are compared, it is also interesting to look at how the same molecules score using related fingerprints. This was studied by calculating all possible pairwise Tanimoto similarities with different fingerprints for the 306 orally active drugs mentioned above and plotting the pairwise similarity values with respect to one fingerprint against the similarity values with respect to another fingerprint. The effect of using feature counts instead of absence/presence bits is exemplified by Figure 5 for the MACCS−MACCSF pair. This plot shows that the two fingerprints correlate, which is expected given that they describe similar structural effects. However, the plot is asymmetric and more points appear below the diagonal than above. Those pairs with similarities between 0.7 and 1 with the MACCS fingerprint are somewhat more spread out using the feature counting version, having similarities between approximately 0.6 and 1. Since, in virtual screening, we are generally interested in those compounds with a high level of similarity to a known active, such an effect may, in part, explain the favorable results for the frequency-based fingerprints in the test cases presented above, particularly at lower sampling levels. The relationship between other related fingerprint pairs is illustrated by descriptive statistical parameters, displayed in Table 4. This table clearly shows that on average the similarity values decrease not only upon introduction of the frequency fingerprints (as discussed above) but also upon introduction of the overlapping feature types. The table also shows an interesting effect: the introduction of overlapping feature types greatly reduces the correlation between the fingerprints, indicating that the modifications to the feature assignments do cause major changes in the fingerprint and are thus likely to result in different types of molecules being picked up in virtual screening. Since the use of overlapping feature types results in an increase in performance, it appears that the new classification scheme describes the action of the groups concerned in the receptor site more accurately than the original type assignments. Such effects can be better studied by looking at individual molecule pairs that might be viewed very differently by different similarity metrics. A few such cases are shown in Figure 6 with more detail given in the captions. These examples illustrate the typical changes to molecular similarity observed upon the introduction of some of the fingerprint design changes described above.

CONCLUSIONS

Because of the success of fingerprint-based virtual screening in day-to-day drug discovery at Neurocrine, we have undertaken some further development of established fingerprints. First, changes were made in the way pharmacophore definitions are handled, allowing different pharmacophore types to overlap and enable partial-type matching. Second, bit string type pharmacophore and structural fingerprints, indicating only the presence and absence of certain features, were altered also to include information on the number of occurrences of those features. Third, the resolution in property description was increased for a property fingerprint.

Although the sets in this paper were all based on screening data from GPCR receptors, the four projects covered contained substantially different types of compounds in terms of both size and chemical features. The consistency of the results across the sets, combined with the fact that the changes were based on chemical intuition rather than derived from data mining, leads the authors to believe that the methods would be transferable to a wide range of problems and thus generally beneficial to virtual screening performance. The two validation sets were designed to approximate two different scenarios in drug discovery, namely, lead hopping (or late lead finding) and lead optimization. Not too surprisingly, it was found that the introduced changes had a greater impact in the lead finding scenario (referred to as the hard case in the manuscript), partly because of the fact that there is more room for improvement. This is in line with our general experience that similarity-based virtual screening is most useful at this stage.

Although some of the described improvements will undoubtedly make a major impact in future virtual screening campaigns, there is further room for improvement. During the process of practical drug discovery, a lot of activity information is obtained, which needs to be fed back to the process. For many fingerprint-based methods, this can be achieved by continually updating the training set. Since current training sets are based on biologically active compounds, these methods essentially learn only from positive data. Although some procedures have been described in the literature for using all available data, if major improvements in practical virtual screening performance are sought, it is critical to find efficient ways to include information from relevant negative activity data into the fingerprints.

ACKNOWLEDGMENT

The assistance of Vu Nguyen in the performance of some of the calculations is gratefully acknowledged.

REFERENCES AND NOTES

(1) Willett, P. Searching techniques for databases of two- and three-dimensional chemical structures. *J. Med. Chem.* **2005**, *48*, 4183−4199.

(2) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177−1185.

(3) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand−receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.

(4) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273−1280.

(5) Flower, D. R. On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.

(6) Shemetulskis, N. E.; Weininger, D.; Blankley, C. J.; Yang, J. J.; Humblet, C. Stigmata: An algorithm to determine structural commonalities in diverse datasets. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862−871.

(7) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem.* **2004**, *2*, 3256−3266.

(8) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: Overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42*, 3251−3264.

(9) McGregor, M. J.; Muskal, S. M. Pharmacophore fingerprinting. 1. Application to QSAR and focused library design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569−574.

(10) Xue L.; Godden J. W.; Stahura F. L.; Bajorath J. Design and evaluation of a molecular fingerprint involving the transformation of property descriptor values into a binary classification scheme. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1151−1157.

(11) Xue, L.; Stahura, F. L.; Godden, J. W.; Bajorath, J. Mini-fingerprints detect similar activity of receptor ligands previously recognized only by three-dimensional pharmacophore-based methods. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 394−401.

(12) Brown, R. D.; Martin, Y. C. Use of structure−activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(13) Baber, J. C.; Shirley, W. A.; Gao, Y.; Feher, M. The use of consensus scoring in ligand-based virtual screening. *J. Chem. Inf. Comput. Sci.* **2006**, *46*, 277−288.

(14) *MOE* (Molecular Operating Environment), version 2005.06; Chemical Computing Group Inc.: Montreal, Canada, 2005.

(15) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T. M.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual screening using protein−ligand docking: Avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 793−806.

(16) Bender, A.; Glen, R. C. A discussion of measures of enrichment in virtual screening: Comparing the information content of descriptors with increasing levels of sophistication. *J. Chem. Inf. Model.* **2005**, *45*, 1369−1375.

(17) Salim, N.; Holliday, J.; Willett, P. Combination of fingerprint-based similarity coefficients using data fusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435−442

(18) Nilakantan, R.; Bauman, N.; Venkataraghavan, R. New method for rapid characterization of molecular shapes: Applications in drug design. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 79−85.

(19) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph kernels for chemical informatics. *Neural Networks* **2005**, *18*, 1093−1110.

(20) *MDDR: MDL Drug Data Report*, 2004.1; Elsevier MDL: San Leandro, CA, 2004.

(21) *SVL Exchange: An SVL code exchange site for the MOE user community*; Chemical Computing Group Inc.: Montreal, Canada.

(22) Zhang, Q.; Muegge, I. Scaffold Hopping Through virtual screening using 2D and 3D similarity descriptors: Ranking, voting, and consensus scoring *J. Med. Chem.* **2006**, *49*, 1536−1548

(23) Brown, R. D.; Martin, Y. C. The information content of 2D and 3D structural descriptors relevant to ligand−receptor binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.

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

CI060155B