

# Novel Application of 2D and 3D-Similarity Searches To Identify Substrates among Cytochrome P450 2C9, 2D6, and 3A4

R. F. Freitas, R. L. Bauab, and C. A. Montanari\*

Grupo de Estudos em Química Medicinal de Produtos Naturais - NEQUIMED-PN, Instituto de Química de São Carlos - Universidade de São Paulo, 13560-970 - São Carlos - SP, Brazil

Received February 26, 2009

Cytochrome P450 (CYP450) is a class of enzymes where the substrate identification is particularly important to know. It would help medicinal chemists to design drugs with lower side effects due to drug–drug interactions and to extensive genetic polymorphism. Herein, we discuss the application of the 2D and 3D-similarity searches in identifying reference structures with higher capacity to retrieve substrates of three important CYP enzymes (CYP2C9, CYP2D6, and CYP3A4). On the basis of the complementarities of multiple reference structures selected by different similarity search methods, we proposed the fusion of their individual Tanimoto scores into a consensus Tanimoto score ( $T_{\text{consensus}}$ ). Using this new score, true positive rates of 63% (CYP2C9) and 81% (CYP2D6) were achieved with false positive rates of 4% for the CYP2C9–CYP2D6 data set. Extended similarity searches were carried out on a validation data set, and the results showed that by using the  $T_{\text{consensus}}$  score, not only the area of a ROC graph increased, but also more substrates were recovered at the beginning of a ranked list.

## INTRODUCTION

Similarity searching methods are widely used in modern drug discovery virtual screening and are based on the similarity principle that states molecules structurally similar are likely to have similar properties.<sup>1</sup> There are a number of reasons for the rapid propagation of similarity searches, which include the less cost-effective computational time, thus allowing searches in huge databases and a pivotal application when there is little (or none) information about the target and only one or two known actives.<sup>2</sup>

Since its introduction, similarity search methods have been focused on the screening of databases to identify new compounds with activity similar to that of a known reference ligand.<sup>3</sup> However, affinity for a target enzyme is only one criterion used to select or discard a compound during the drug discovery process and constitutes the pharmacodynamic phase of a drug. The other key requirement for a safer drug is a suitable pharmacokinetic profile, which incorporates the study of ADME/Tox (absorption, distribution, metabolism, excretion, along with toxicity) properties. These studies have become an essential task to reduce the attrition rate at the late stages of the drug development process. In 1991, both poor bioavailability and pharmacokinetics were the properties responsible for 40% of all attrition rates in the drug discovery pipeline. This number dramatically dropped to 10% in 2000<sup>4</sup> after the inclusion of ADME/Tox studies at earlier stages in the pipeline.

Drug metabolism is a crucial pharmacokinetic property where substrate identification is particularly important to know. It is the phase of biochemical transformation of the drug, and it is traditionally divided into phase I and phase II processes. The first involves the modification of a functional

group by oxidation, reduction, or hydrolysis. The second is responsible for the conjugation of the phase I metabolite with an endogenous molecule such as glucuronic acid.<sup>5</sup> Cytochrome P450 (CYP) enzymes play a key role in phase I metabolism, where the goal is to convert the drug into a polar form by adding an ionizable group, thus making them more water-soluble and more readily to be excreted.

In this regard, CYP inhibition is a major safety problem. Very often, two or more drugs are coadministered to a patient during the treatment of a disease, and the individual compounds may compete to be metabolized by the same enzyme. This leads to unintended effects of drug–drug interactions where a drug inhibits the metabolism, causing an increase in the plasma concentration of a second drug, which can lead to an increase in the toxic side effects.<sup>6,7</sup> That is, measurement of CYP inhibition during discovery provides early warning of potential safety issues. This means CYP substrate selectivity is sometimes much less of an issue and in fact there may be potential advantages in having a drug that is a substrate for more than one CYP because if there is a problem (inhibition or individual variation) with one CYP another one can take over. In addition, polymorphisms (result of genetic mutations) affect the activity of a number of drug-metabolizing enzymes such as CYP2C19, CYP2D6, and CYP1A2. For example, the polymorphic enzyme CYP2D6 is absent in 5–10% of Caucasians and gives rise to poor metabolizers that have low or lack of activity.<sup>8</sup> Hence, it would be highly beneficial for the discovery and development of safer drugs to introduce *in silico* methods to predict, as early as possible, the structural features that imprint substrate identification. Additionally, because the drug molecule can be metabolized to either reactive or chemically stable metabolites, leading, respectively, to drug-induced toxicity or enhanced pharmacology, such methods would be of utmost importance.

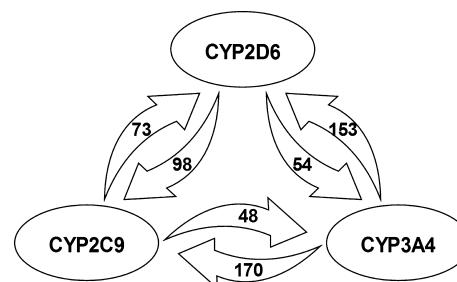
\* Corresponding author phone: 55-16-3373-9986; fax: 55-16-3373-9985; e-mail: montana@iqsc.usp.br.

Computational work has hitherto experienced the immediate possibility to predict substrate selectivity mainly focused on structure-based methods such as docking using homology models and X-ray structures of the CYP enzymes.<sup>9,10</sup> Ligand-based approaches such as pharmacophore<sup>11,12</sup> and QSAR<sup>13,14</sup> analyses have also been extensively employed. Based upon the need to further the knowledge in this field, the present work introduces a novel approach that expands the scope of similarity searching methods from the classical application for a single target to deal with the challenging task of explaining the substrate identification of relevant CYP enzymes. For a statistically significant database of 596 substrates of the three most important CYP enzymes (CYP2C9, CYP2D6, and CYP3A4), 2D and 3D-similarity searches were used in the determination of the CYP enzyme predominantly responsible for the metabolism of a compound. These enzymes are highly promiscuous, which means that they are able to metabolize a broad range of chemically diverse substrates and many times more than one isoform metabolizes the same substrate, which has previously been mentioned to be sometimes advantageous where selectivity is not necessarily an issue. Herein, we show a successful application of a potentially powerful method to hot spot reference structures that enables the identification and separation among substrates and nonsubstrates for a particular CYP isoform.

## METHODS

**Data Sets.** Drugs that are substrates of CYP2C9, CYP2D6, and CYP3A4 were mainly collected from the previous work of Rendic,<sup>15</sup> Terfloth,<sup>16</sup> and Yap.<sup>17</sup> As we carried out pairwise analysis between these three cytochromes (2C9 versus 2D6, 2C9 versus 3A4, and 2D6 versus 3A4), for each compound, a search in the literature was carried out to confirm if it was entirely metabolized by the CYP enzyme according to the statement in the three papers above. If a compound is metabolized by more than one CYP450, it was classified as being a substrate of the isoform responsible for the main route of its metabolism. To exemplify our strategy to build strong relationships between the desired CYP and the substrate to be metabolized, compounds were hand-selected to allow the outstanding achievement of identifying the CYP for that metabolism. Albeit this is time-consuming, it ensures the high quality of the data. With this protocol, we were able to construct meaningful pairwise substrate sets. For example, in Figure 1, 73 compounds could be used to analyze CYP2C9 over CYP2D6, and 98 for CYP2D6 over CYP2C9 (opposite direction).

The generated data sets are composed of 596 compounds, out of which 417 are unique. The entire set is composed by 121 CYP2C9 substrates (20%), 151 CYP2D6 substrates (25%), and 324 CYP3A4 substrates (55%). From the analysis of Figure 1 and the previous statistics, the data sets CYP2C9–CYP3A4 and CYP2D6–CYP3A4 are very unbalanced. This observation agrees with the fact that more than one-half of all marketed drugs are metabolized by the CYP3A4 enzyme. The whole number of compounds differs from the number of unique compounds because a substrate could be present in more than one pairwise analysis. For example, the drug aceclofenac



**Figure 1.** Diagram representing CYPs compound set. Each arrow represents a set, and its direction defines the CYP for that metabolism. For example, in the CYP2C9–CYP2D6 data set, we have 171 compounds. Out of these, 73 are CYP2C9 substrates, and 98 are CYP2D6 substrates.

was considered a substrate of CYP2C9 in both pairwise analysis, CYP2C9–CYP2D6 and CYP2C9–CYP3A4. The reported CYP substrate benchmark will be made freely available to support compound identification analyses using other computational approaches.

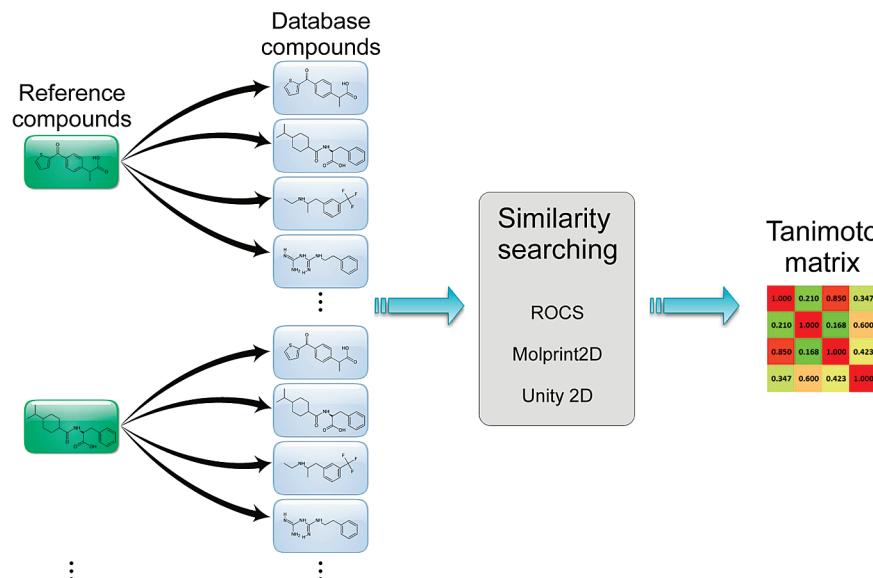
**Molecular Similarity.** On the basis of the similar property principle,<sup>1</sup> which states that structural similar molecules are more likely to have similar properties, we carried out an all-against-all similarity analysis for every compound in our data set to identify the desired CYP substrate. Systematic pairwise similarity calculations were fulfilled to evaluate the structural diversity distribution of each pair of enzymes. The similarity between the compounds was measured on the basis of the Tanimoto coefficient ( $T_c$ ). Figure 2 shows the workflow for the approach adopted in this work. Each compound was used as reference compound, and a similarity value was calculated to all compounds in the database. This leads to a square similarity Tanimoto matrix with the number of rows and columns equal to the number of compounds (Figure 2).

In the present work, we investigate the performance of 2D and 3D-similarity searches for their ability to distinguish compounds between closely related targets. For 2D-similarity searching, two 2D fingerprints were selected that are calculated from 2D molecular graphs: Molprint2D and Unity 2D fingerprint.<sup>18</sup> In addition, ROCS (rapid overlay of chemical structures) was used to perform 3D-similarity searches. Tanimoto matrices were generated using the publicly available Matrix2png software.<sup>19</sup>

**Quantification of Performance.** The evaluation of the different methods to discriminate between substrates and nonsubstrates was carried out by using the area under the curve (AUC) of a receiver operating characteristic (ROC) curve. Plotting a ROC curve consists of the determination of the sensitivity ( $Se$ ) and the specificity ( $Sp$ ) at every possible score threshold.<sup>29</sup> The first value ( $Se$ ) describes the ratio of the number of true actives that are selected by the method to the number of all actives in the database:

$$Se = \frac{N_{\text{selected actives}}}{N_{\text{total actives}}} = \frac{TP}{TP + TN}$$

The second value ( $Sp$ ) describes the ratio of the number of inactives that are discarded by the method to the number of all inactive molecules included in the databases:



**Figure 2.** Workflow representation of the pairwise similarity calculations.

$$Se = \frac{N_{\text{discarded inactives}}}{N_{\text{total inactives}}} = \frac{TP}{TP + TN}$$

Both of these parameters can vary between 0 and 1. Usually, the greater is the AUC, the more effective the virtual screening workflow is in discriminating active from inactive compounds.<sup>29</sup>

## RESULTS AND DISCUSSION

The analysis of some physicochemical properties of the compounds studied here shows that they have similar values for almost all properties. The molecules belonging to the CYP2C9 and CYP3A4 sets are the ones that present more related properties, with the molecular weight (MW) being a little large in the CYP3A4 set. Despite that the CYP2D6 set also displays many comparable properties with the other two data sets, the mean molecular weight and mean polar surface area (PSA) have the smallest values of the whole data set, which indicate that the CYP2D6 compounds are smaller than that of the other two enzymes. The constitution of data sets has a pronounced effect on the efficacy of a virtual screening test.<sup>20</sup> For instance, in a recent work of Jain et al., the performance of 3D virtual screening methods was comparable to a simple 1D method (molecular weight, log *P*, number of hydrogen-bond donors, number of hydrogen-bond acceptors, and number of rotatable bonds).<sup>21</sup> This result indicates that the active compounds are very dissimilar from the inactive set, because a simple 1D method is very efficient in the separation of these two sets. Therefore, the good results found in many publications in the area of retrospective virtual screening are purely due to differences in simple properties between the actives and the inactives. Next, an inactive set should be as similar as possible to the active compounds to obtain a trust indication of the utility of the virtual screening method. Therefore, our virtual library is appropriate to assess the usefulness of the tools employed in the present work because there is no statistical difference within the simple 1D molecular properties (Table 1).

**Compound Diversity.** The intraset structural diversity of the CYP2C9–CYP2D6 data set was evaluated using sys-

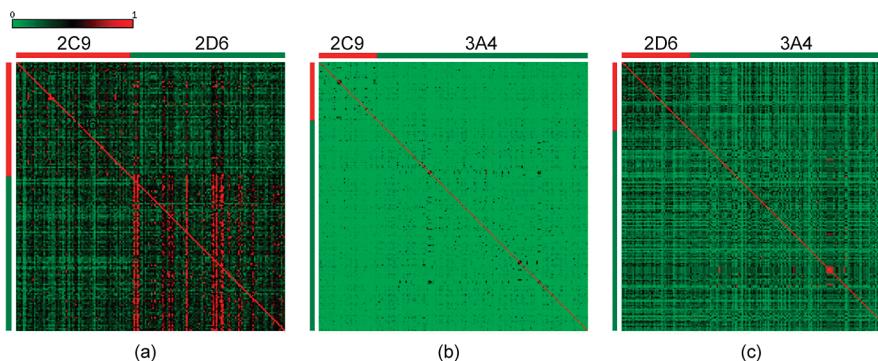
**Table 1.** Mean and Standard Deviation of Some 1D Properties of the Substrates of Our Database

	CYP2C9		CYP2D6		CYP3A4	
	mean	std. dev.	mean	std. dev.	mean	std. dev.
HBA <sup>a</sup>	5	2	4	2	5	3
HBD <sup>b</sup>	1	1	2	1	2	1
MW <sup>c</sup>	321	88	303	77	369	114
RB <sup>d</sup>	5	3	5	3	5	3
log <i>P</i> <sup>e</sup>	3	2	3	2	3	2
PSA <sup>f</sup>	69	31	44	23	70	35

<sup>a</sup> Hydrogen-bonding acceptor. <sup>b</sup> Hydrogen-bonding donor. <sup>c</sup> Molecular weight. <sup>d</sup> Rotatable bonds. <sup>e</sup> Partition coefficient. <sup>f</sup> Polar surface area.

tematic pairwise similarity comparisons. Because each compound was used as reference compound and a similarity value was calculated to all compounds in the database, the result was a square similarity Tanimoto matrix with the number of rows and columns equal to the number of compounds (Figure 3a–c). As we would expect, there is a diagonal line (red), where the similarity is maximum, which means that the similarity was measured using the same structure, or, in other words, the reference and the database compound are equals. We can observe regions presenting high similarities outside that diagonal, indicating that many substrates are also very similar to other ones.

For the 3D-similarity search (ROCS), the compound structural diversity, quantified by the Tanimoto coefficient, considerably differs depending on the level of information used to score the compounds and on the method used. Analyzing the results of the ROCS method after the inclusion of chemical features (color score) is possible to observe that the CYP2D6 substrates present a high intrasimilarity, as observed by the large right-bottom cluster in the matrix (Figure 3a), meaning that these substrates have a great homogeneity in their structures. With regard to the CYP2C9 substrates, it is possible to verify some disperse points in the matrix presenting a high similarity, but in general no significant cluster is apparent, indicating that the substrates of this isoform display a higher chemical diversity if compared to the substrates from CYP2D6 isoform. The



**Figure 3.** Tanimoto matrices for pairwise comparisons of the three datasets: (a) 2C9–2D6/ROCS-color; (b) 2C9–3A4/MOLPRINT2D; (c) 2D6–3A4/Unity 2D. Similarity values were assessed using the Tanimoto coefficient ( $T_c$ ). For  $T_c$  values, continuous color-coding is used with increasing similarity from green to red (black indicates  $T_c = 0.5$ ). In each matrix, the position of the compounds for a particular substrate selectivity set is marked by color bars.

comparison of molecules on both shape and chemical complementarity was effective to produce a cluster in the Tanimoto matrix for the CYP2D6 substrates (Figure S1).

The similarity searching using fingerprints, Molprint2D and Unity 2D, provided results similar to those with ROCS-color. Using Molprint2D fingerprint, the compounds appear to have the most diverse structures, as observed by their extremely low pairwise similarity in the Tanimoto matrix (Figure S1). However, inside each class, the similarity is higher, and it is possible to see the two clusters, one for the CYP2C9 substrates and the other for the CYP2D6 substrates. The Unity 2D fingerprint shows a behavior similar to the one observed for Molprint2D, with the CYP2D6 substrates presenting high pairwise similarities (Figure S1).

For the CYP2C9–CYP3A4 data set, Molprint2D provided a clear cluster in the diversity matrix at the top left of the matrix, which corresponds to the CYP2C9 substrates (Figure 3b). Using ROCS, independently of the kind of score used, it was not possible to identify a separation between the substrates of the CYP2C9 and CYP3A4 enzymes (Figure S2). Unity 2D shows an analogous behavior to the viewed for ROCS, where no cluster is observed in the Tanimoto matrix (Figure S2).

The systematic pairwise similarity comparison of the CYP2D6–CYP3A4 data set using the Unity 2D fingerprint shows a top-left cluster including all CYP2D6 substrates (Figure 3c). The remaining areas of the matrix presented low pairwise similarity among the structures, with the absence of any cluster. Molprint2D and ROCS show a Tanimoto matrix similar to that viewed for Unity 2D, a top-left cluster corresponding to the CYP2D6 substrates, and an absence of any cluster in the CYP3A4 area (Figure S3).

It is noteworthy that no method was able to produce a significant cluster in the Tanimoto matrix for the CYP3A4 substrates in the data sets where they were present, with the exception of some very disperse small areas of high intrasimilarity. These results demonstrate that its substrates have a huge chemical diversity, and this characteristic is in agreement with the broad range of substrate specificity displayed by CYP3A4 enzyme, which metabolizes nearly 50% of the marketed drugs.<sup>22</sup>

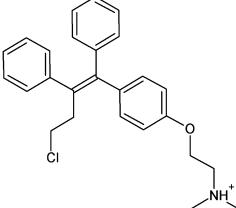
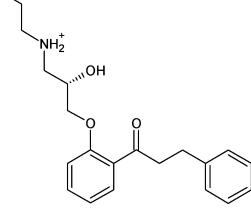
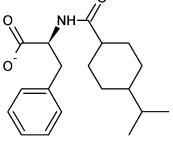
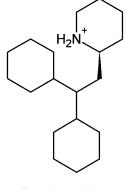
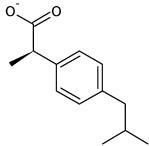
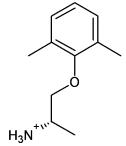
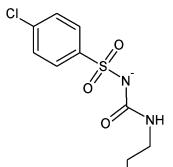
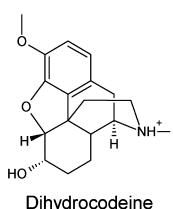
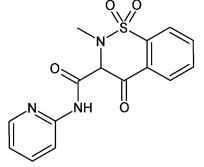
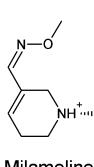
**ROC-AUC Analysis.** The main objective of the present work was to answer the following question: is it possible to identify reference structures that provide a good separation between the substrates of two CYP isoforms using 2D and

3D-similarity searches? To answer this question, we carried out pairwise similarity searches using each compound as a reference and database compound. Next, the performance of the different similarity searches was assessed by measuring the area under the receiver operator characteristic (ROC) curve for every compound used as a reference. Such graphs show the performance of a given tool when the screening across the entire database is examined. ROC curve has been shown to be a powerful technique for investigating the ability of retrospective virtual screening methods to discriminate between active and inactive compounds.<sup>35</sup>

Despite the high chemical diversity exhibited by the CYP2C9 substrates, all methods were able to identify reference structures that allowed the identification of substrates and nonsubstrates for this enzyme. Table 2 shows the chemical structures of the substrates that were more efficient to distinguish the substrates of the two isoforms (CYP2C9 and CYP2D6) according to a given similarity search. The one where the reference structure gave the best performance has its AUC value highlighted in bold. For example, the substrate toremifene is the one that has the highest AUC value among all CYP2C9 substrates using the ROCS-shape. Just for comparison, the AUC values using other similarity searches are also displayed. The best performance in the separation of CYP2C9 and CYP2D6 substrates was provided by ROCS-color (0.807), ROCS-combo (0.755), and Unity 2D (0.707), using nateglinide, ibuprofen, and piroxicam, respectively, as reference structures. These AUC values mean that the score of a randomly selected substrate is higher than a randomly selected nonsubstrate 7 times out of 10. Also, it is clear from the analysis of Table 2 that the results from ROCS are better when using some kind of chemical typing rather than just shape, because the AUC goes from 0.566 (ROCS-shape) to 0.807 (ROCS-color). The substrates toremifene (ROCS-shape) and chlorpropamide (Molprint2D) presented only marginal AUC values.

It is also possible to see from Table 2 the reference structures that were more efficient to separate substrates and nonsubstrates of the CYP2D6 isoform. Using the substrate perhexiline as reference compound, we were able to separate CYP2D6 substrates from CYP2C9 with an impressive AUC value of 0.912. In addition, the AUC values for three reference structures are higher than 0.850.

**Table 2.** Best Reference Structures and the Respective AUC Values (Bold) That Identified Substrates of the Enzymes CYP2C9 and CYP2D6<sup>a</sup>

CYP2C9		CYP2D6	
Structure	Method AUC	Structure	Method AUC
	<b>ROCS-shape</b> <b>0.566</b> ROCS-color 0.438 ROCS-combo 0.459 Molprint2D 0.371 Unity 2D 0.250		<b>ROCS-shape</b> <b>0.581</b> ROCS-color 0.667 ROCS-combo 0.689 Molprint2D 0.625 Unity 2D 0.684
	ROCS-shape 0.357 <b>ROCS-color</b> <b>0.807</b> ROCS-combo 0.639 Molprint2D 0.561 Unity 2D 0.372		ROCS-shape 0.398 <b>ROCS-color</b> <b>0.912</b> ROCS-combo 0.822 Molprint2D 0.646 Unity 2D 0.849
	ROCS-shape 0.514 ROCS-color 0.721 <b>ROCS-combo</b> <b>0.755</b> Molprint2D 0.547 Unity 2D 0.260		ROCS-shape 0.550 ROCS-color 0.901 <b>ROCS-combo</b> <b>0.893</b> Molprint2D 0.572 Unity 2D 0.716
	ROCS-shape 0.456 ROCS-color 0.617 ROCS-combo 0.560 <b>Molprint2D</b> <b>0.630</b> Unity 2D 0.512		ROCS-shape 0.513 ROCS-color 0.719 ROCS-combo 0.615 <b>Molprint2D</b> <b>0.795</b> Unity 2D 0.693
	ROCS-shape 0.539 ROCS-color 0.593 ROCS-combo 0.562 Molprint2D 0.625 <b>Unity 2D</b> <b>0.707</b>		ROCS-shape 0.439 ROCS-color 0.844 ROCS-combo 0.834 Molprint2D 0.624 <b>Unity 2D</b> <b>0.857</b>

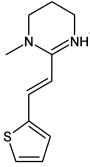
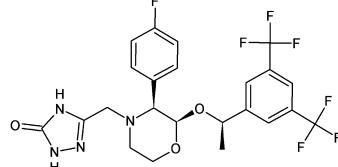
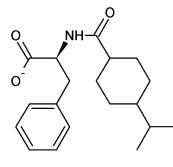
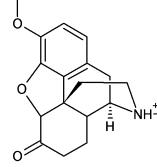
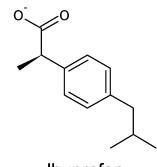
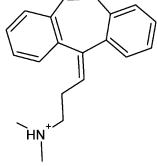
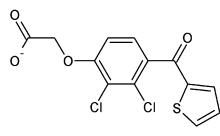
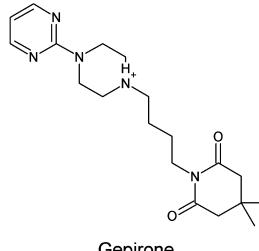
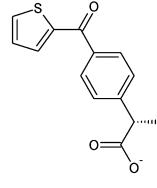
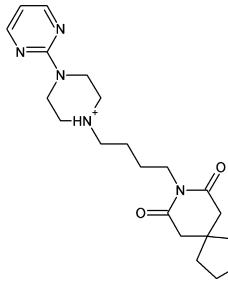
<sup>a</sup> For comparison, the AUC values of all methods are also displayed.

Another interesting result shown in Table 2 is that all reference structures of the CYP2D6 enzyme display a protonated nitrogen in their structures. Our results are in agreement with other works where a usual characteristic of the majority of CYP2D6 substrates is the presence of a basic nitrogen atom and an aromatic ring.<sup>24,25</sup> In addition, CYP2C9 substrates are usually weak acids with multiple aromatic

rings. Looking at Table 2, we see that three reference structures meet this structural pattern.

The reference structures that provided the best separation between the substrates of the CYP2C9 and CYP3A4 enzymes are displayed in Table 3. Direct information that emerges from the analysis of this table is that the substrates nateglinide and ibuprofen were again the best ones to separate CYP2C9

**Table 3.** Best Reference Structures and the Respective AUC Values (Bold) That Provided a Separation between the Substrates of the Enzymes CYP2C9 and CYP3A4<sup>a</sup>

CYP2C9		CYP3A4	
Structure	Method AUC	Structure	Method AUC
	<b>ROCS-shape</b> <b>0.684</b> ROCS-color 0.318 ROCS-combo 0.469 Molprint2D 0.538 Unity 2D 0.431		<b>ROCS-shape</b> <b>0.668</b> ROCS-color 0.557 ROCS-combo 0.643 Molprint2D 0.502 Unity 2D 0.623
	ROCS-shape 0.478 <b>ROCS-color</b> <b>0.742</b> ROCS-combo 0.690 Molprint2D 0.561 Unity 2D 0.395		ROCS-shape 0.412 <b>ROCS-color</b> <b>0.755</b> ROCS-combo 0.590 Molprint2D 0.607 Unity 2D 0.618
	ROCS-shape 0.650 ROCS-color 0.640 <b>ROCS-combo</b> <b>0.739</b> Molprint2D 0.623 Unity 2D 0.505		ROCS-shape 0.424 ROCS-color 0.732 <b>ROCS-combo</b> <b>0.683</b> Molprint2D 0.537 Unity 2D 0.538
	ROCS-shape 0.675 ROCS-color 0.552 ROCS-combo 0.694 <b>Molprint2D</b> <b>0.669</b> Unity 2D 0.586		ROCS-shape 0.425 ROCS-color 0.671 ROCS-combo 0.532 <b>Molprint2D</b> <b>0.689</b> Unity 2D 0.696
	ROCS-shape 0.645 ROCS-color 0.646 ROCS-combo 0.729 Molprint2D 0.663 <b>Unity 2D</b> <b>0.624</b>		ROCS-shape 0.464 ROCS-color 0.671 ROCS-combo 0.559 Molprint2D 0.662 <b>Unity 2D</b> <b>0.707</b>

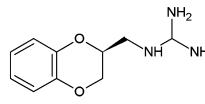
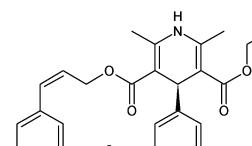
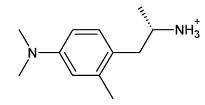
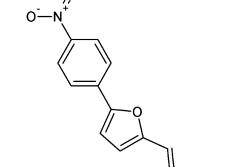
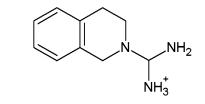
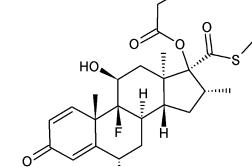
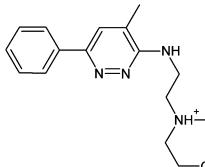
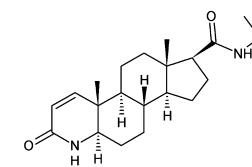
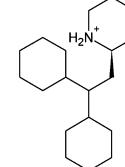
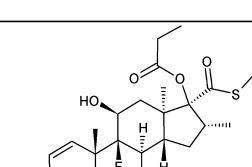
<sup>a</sup> For comparison, the AUC values of all methods are also displayed.

substrates from those of the CYP3A4 isoform, indicating that they are typical structures of the CYP2C9 substrates. Another intriguing result is that three of five reference structures of the CYP2C9 enzyme belong to the class of nonsteroidal anti-inflammatory drugs (NSAIDs): ibuprofen, suprofen, and tielinic acid. The substrate nateglide belongs to the megl-

tinide class of blood glucose-lowering drugs, but as the other three substrates, it also presents the usual structure of the CYP2C9 substrates: a carboxylate group and at least one aromatic ring.

We can see that all reference structures of the CYP3A4 enzyme displayed an AUC value close to 0.700, with

**Table 4.** Best Reference Structures and the Respective AUC Values (Bold) That Provided a Separation between the Substrates of the Enzymes CYP2D6 and CYP3A4<sup>a</sup>

CYP2D6		CYP3A4	
Structure	Method AUC	Structure	Method AUC
	<b>ROCS-shape 0.741</b> ROCS-color 0.747 ROCS-combo 0.844 Molprint2D 0.730 Unity 2D 0.754		<b>ROCS-shape 0.673</b> ROCS-color 0.486 ROCS-combo 0.596 Molprint2D 0.434 Unity 2D 0.666
	ROCS-shape 0.631 <b>ROCS-color 0.806</b> ROCS-combo 0.851 Molprint2D 0.668 Unity 2D 0.772		ROCS-shape 0.415 <b>ROCS-color 0.822</b> ROCS-combo 0.603 Molprint2D 0.539 Unity 2D 0.677
	ROCS-shape 0.711 ROCS-color 0.783 <b>ROCS-combo 0.880</b> Molprint2D 0.680 Unity 2D 0.826		ROCS-shape 0.615 ROCS-color 0.668 <b>ROCS-combo 0.698</b> Molprint2D 0.573 Unity 2D 0.784
	ROCS-shape 0.568 ROCS-color 0.683 ROCS-combo 0.671 <b>Molprint2D 0.760</b> Unity 2D 0.615		ROCS-shape 0.519 ROCS-color 0.608 ROCS-combo 0.566 <b>Molprint2D 0.685</b> Unity 2D 0.485
	ROCS-shape 0.541 ROCS-color 0.769 ROCS-combo 0.761 Molprint2D 0.592 <b>Unity 2D 0.843</b>		ROCS-shape 0.615 ROCS-color 0.668 ROCS-combo 0.698 Molprint2D 0.573 <b>Unity 2D 0.784</b>

<sup>a</sup> For comparison, the AUC values of all methods are also displayed.

hydrocodone showing the highest value (0.755). The same observation emerges for the reference structures of the CYP2C9 enzyme. These results reveal the difficulty to separate the substrates of these two isoforms. These can be explained by the fact that these two isoforms are able to metabolize a wide range of chemically diverse substrates. It is quite common that the substrates overlap between these CYP450 enzymes where more than one enzyme metabolizes the same substrate.

From the analysis of Table 4, we can see the reference structures selected by each method to separate the CYP3A4

substrates from those of CYP2D6 enzyme. Of the five reference structures, two (fluticasone and finasteride) present the lipophilic steroid scaffold, characterized by a terpenoid lipid with a carbon skeleton with four fused rings, arranged in a 6–6–6–5 fashion. Actually, the substrate fluticasone was selected by ROCS-combo and Unity 2D fingerprint as the best reference structure. This result agrees with the preference of CYP3A4 to catalyze the oxidation of lipophilic neutral or basic compounds.

The presence of basic nitrogen in the CYP2D6 substrates is evident if we look at Table 4, which shows the

best reference structures to separate the substrates of the CYP2D6 and CYP3A4 enzymes. In addition, three of these structures show AUC values higher than 0.800, indicating that a higher score is assigned 8 times out of 10 to a randomly selected substrate more than that of a randomly selected nonsubstrate. Too, the substrate perhexiline is found to be the best reference structure using ROCS-color in the analysis of the CYP2C9–CYP2D6 data set. It was selected again using Unity 2D fingerprint, validating its structure as an efficient reference structure to identify CYP2D6 substrates.

Before concluding, some further considerations on methods performance and discrimination among substrates and nonsubstrates have to be provided. Clearly, shape score is the poorest of the three available in ROCS (shape, color, and combo score). There is a sensible improvement in ROCS results when we use color and combo score. The two fingerprints used in this work (Molprint2D and Unity 2D) perform as well as ROCS-shape score and in some cases better, but with the advantage that they are less computationally expensive. Several papers have also found that for various target classes, 2D-similarity searches using fingerprints have given comparable performance or even superior results to the 3D methods.<sup>2,26,27</sup> A possible explanation might be that the connection table of a molecule encodes implicit information about the structure of a molecule that is lost in the 3D-methods, which ignores bond topology in favor of atom positions.<sup>2</sup> Comparing just the two fingerprints, we see that Unity 2D works better than Molprint2D, because in five out of six reference structures the area under the ROC curve is higher when using the former fingerprint. They only show comparable performance for the CYP2C9–CYP3A4 data set. The other consideration is about the difficulty to discriminate between substrates and nonsubstrates for the three data sets. According to the AUC values achieved by each reference structure, we can see that the most challenging is the CYP2C9–CYP3A4 data set, because it is the one that is constituted by substrates of the two most chemically diverse substrates. At the other extreme, we have the CYP2C9–CYP2D6 data set. Identifying CYP2D6 substrates among the ones from CYP2C9 is not a difficult task because both enzymes are often represented by distinctly compounds that occupy different chemical spaces: CYP2C9 substrates are usually weak acids, and the CYP2D6 substrates have a basic nitrogen atom in its structures. Finally, the CYP2D6–CYP3A4 is at an intermediate position.

**Consensus Similarity.** In the light of the comparable efficiency of 2D and 3D-similarity searches, we have noted that despite that some method could be more robust to identify substrates on the whole, it is practically unfeasible for only one method to distinguish unambiguously all of the substrates of enzymes highly promiscuous like the cytochrome P450. Clearly, each method can find some structures that all other methods would miss. Therefore, we think the union of different similarity searching methods and multiple reference structures is a good strategy to retrieve as many substrates as possible for challenging enzymes like the CYPs. Actually, similar strategies were proposed in the works of Muresan et al.<sup>28</sup> and Willett et al.<sup>29</sup> The former carried out a multifingerprint selection by merging the compounds coming from a ranked list of selected targets corresponding to all fingerprints. They claimed that a multifingerprint

approach could be an efficient tool to balance the strengths and weaknesses of various fingerprints.<sup>28</sup> In the work of Willett et al., they merged individual fingerprints of a set of 10 reference structures into a single combined fingerprint. This resulted in an increase of over two-thirds in the numbers of actives retrieved.<sup>29</sup> In another recent work, Gasteiger et al. found that the use of similarity search with subsequent data fusion produced up to 16% better BEDROC scores than the novelty detection with self-organizing maps.<sup>30</sup>

To investigate our hypothesis, we took the best three reference structures for each data set and combined their Tanimoto score in a new score called consensus Tanimoto score ( $T_{\text{consensus}}$ ). The scores of each method were scaled to a number between 0 and 1 using the following formula:

$$T_{\text{scaled}} = (T - T_{\min}) / (T_{\max} - T_{\min}) \quad (1)$$

Here,  $T_{\text{scaled}}$  is the scaled Tanimoto values of the  $i_{\text{score}}$  (for example, the Tanimoto values of the ROCS-color),  $T$  is the raw value of the  $i_{\text{score}}$ ,  $T_{\min}$  is the smallest Tanimoto value for a given method, and  $T_{\max}$  is the largest Tanimoto value for a given method. To obtain the consensus Tanimoto score,  $T_{\text{consensus}}$ , we carried out the sum of the three  $T_{\text{scaled}}$  selected, and this value was divided by 3 according to eq 2.

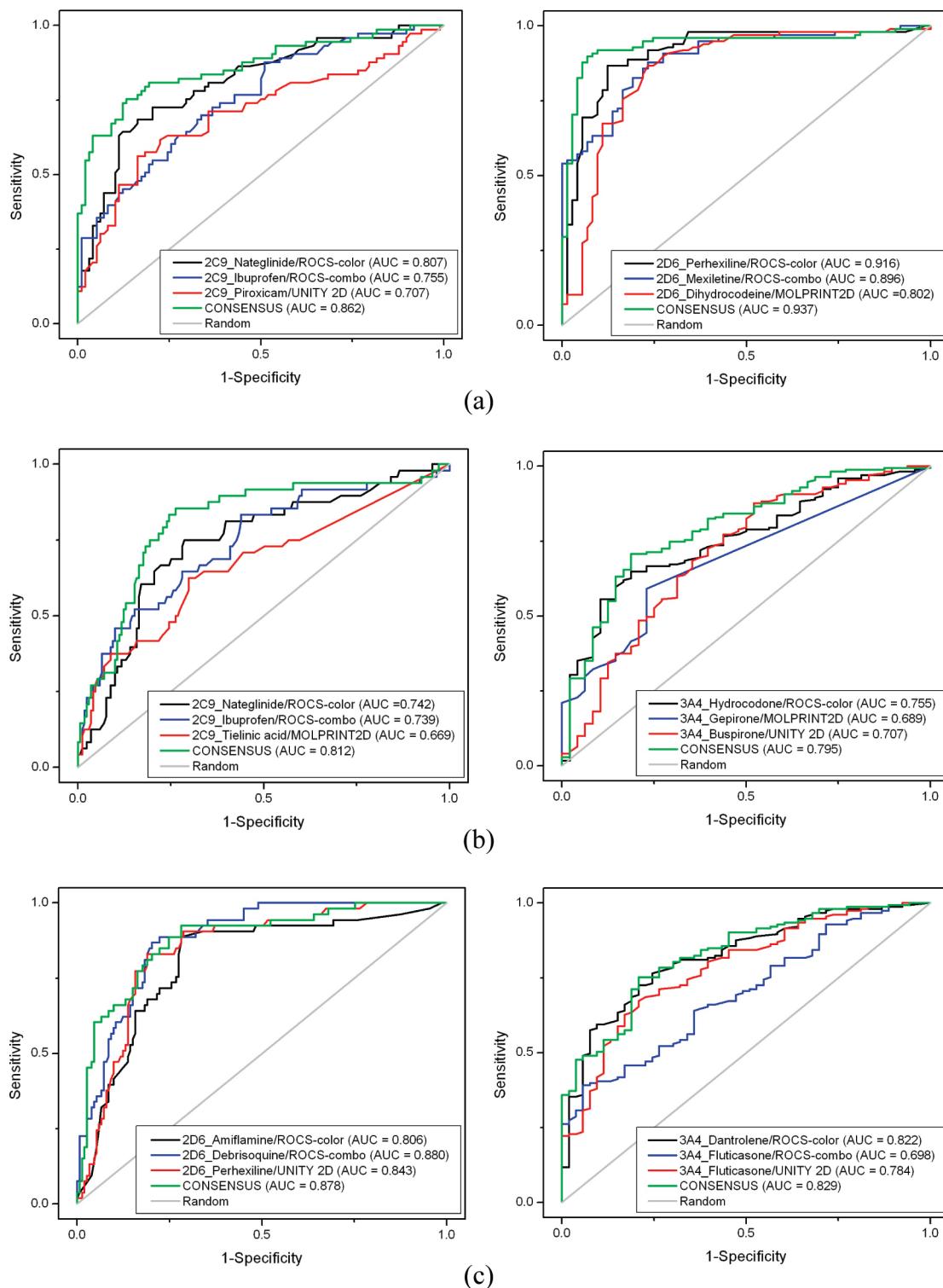
$$T_{\text{consensus}} = \sum_{i_{\text{score}}} T_{\text{scaled}} / 3 \quad (2)$$

With this scaling process, all of the Tanimoto scores can be simultaneously employed in consensus similarity analysis, regardless of the differences between them, such as, for example, the ROCS-combo score, which is a combination of shape and color score, with has a maximum value equal to 2.

By the analysis of Figure 4, we can see the use of the  $T_{\text{consensus}}$  score improves the AUC values for almost all data sets. Only one exception was observed in the case of the CYP2D6–CYP3A4 data set (Figure 4c), where the performance of the new  $T_{\text{consensus}}$  score (AUC = 0.878) was similar to that using just the compound debrisoquine as a reference structure (AUC = 0.880).

ROC curves have many advantages over the more traditional graphical method (enrichment curves). First, they do not depend on the ratio of the screened set of molecules. Second, they provide the entire spectrum of sensitivity/specificity pairs, the only one that supplies a complete picture of test accuracy reporting the dual aspect of any test, the ability to select active compounds and discard inactive ones.<sup>29</sup> However, ROC curves have an important drawback: they are not able to deal with the early recognition problem.<sup>31,32</sup> For instance, consider the case where two ROC curves have the same AUC of exactly 0.5. In the first curve, one-half of the actives are retrieved at the very beginning of the rank-ordered list and the other one-half at the end, while in the second curve, all of the actives are retrieved in the middle of the list. Clearly, the first curve, in terms of the early recognition problem, is better than the second curve. Because the retrieval of active structures on the top ranked list is the main aim of a virtual screening test, saving us the trouble of actually screening all of the compounds, it is entirely reasonable to prefer good “early” behavior.<sup>30,31</sup>

On the basis of the previous paragraph, we consider the observation that in five out six curves, our strategy was



**Figure 4.** ROC curves comparing the performance between isolated reference structures with the consensus of three reference structures: (a) CYP2C9–CYP2D6; (b) CYP2C9–CYP3A4; and (c) CYP2D6–CYP3A4.

responsible for the identification of more substrates at the beginning of the ROC curve, even more important than the improvement of the area under the curve. In the CYP2C9–CYP2D6 data set, using the  $T_{\text{consensus}}$  score, a high sensitivity (63% and 81% for CYP2C9 and CYP2D6, respectively) was achieved at a low 1-specificity of only 4% (Figure 4a). The same sensitivity in the CYP2C9–CYP3A4 data set was achieved at a higher 1-specificity (14–15%) (Figure 4b). For the CYP2D6–CYP3A4 data set, a sensitivity and 1-specific-

ity of 60% and 4.6%, respectively, were obtained using the  $T_{\text{consensus}}$  score for the separation of CYP2D6 substrates (Figure 4c). The only exception mentioned above was found in the separation of CYP3A4 substrates that showed a sensitivity and 1-specificity equal to 60% and 19%, respectively. These results reinforce the challenging task of identifying CYP3A4 substrates using similarity search methods, mainly because of its large and varied “menu” of structures.

**Table 5.** Results from a Similarity Search Carried Out in an External Set of CYP Substrates Using the Best Reference Structures and the  $T_{\text{consensus}}$  Score (Consensus)

CYP2C9–CYP2D6 ( $n = 53$ )					
CYP2C9			CYP2D6		
reference structure	method	AUC	reference structure	method	AUC
nateglinide	ROCS-color	0.865	perhexiline mexiletine milameline consensus	ROCS-color	0.944
ibuprofen	ROCS-combo	0.762		ROCS-combo	0.917
piroxicam	Unity 2D	0.760		Unity 2D	0.790
consensus		0.913			0.955
CYP2C9–CYP3A4 ( $n = 70$ )					
CYP2C9			CYP3A4		
reference structure	method	AUC	reference structure	method	AUC
nateglinide	ROCS-color	0.559	hydrocodone	ROCS-color	0.609
ibuprofen	ROCS-combo	0.797	gepirone	Molprint2D	0.565
tielinic acid	Molprint2D	0.560	buspirone	Unity 2D	0.620
consensus		0.781	consensus		0.652
CYP2D6–CYP3A4 ( $n = 79$ )					
CYP2D6			CYP3A4		
reference structure	method	AUC	reference structure	method	AUC
amiflamine	ROCS-color	0.858	dantrolene	ROCS-color	0.714
debrisoquine	ROCS-combo	0.861	fluticasone	ROCS-combo	0.664
perhexiline	Unity 2D	0.830	fluticasone	Unity 2D	0.756
consensus		0.929	consensus		0.814

**Validation.** Our strategy was further validated with an external data set of 99 substrates (21 2C9 substrates, 31 2D6 substrates, and 47 3A4 substrates). These compounds are distinct from those used in the model development and were collected from various sources.<sup>33,34</sup>

Analyzing Table 5, we can see that the reference structures previously selected were able to correctly classify the new substrates with a higher efficiency. In addition, the new  $T_{\text{consensus}}$  score allowed us to achieve impressive AUC values ranging from 0.652 to 0.955. This performance was similar to, and in some cases better than, the AUC values displayed in Figure 4. The poorest result, but not totally unacceptable, comes from the CYP3A4 substrates. It is noticeable that this bad result is tightly associated with the diversity of the substrates of this enzyme. The larger is the chemical diversity of a set of compounds, the more difficult it is to find a reference structure able to retrieve the majority of the compounds. Nevertheless, the use of the  $T_{\text{consensus}}$  score gave rise to a moderate improvement in the AUC values when compared to the search using a single reference structure.

For the validation data set, as observed during the model development, the use of the  $T_{\text{consensus}}$  score not only increased the area of a ROC graph, but also increased the recovery of the substrates at the beginning of the ranked list. Using the  $T_{\text{consensus}}$  score, the sensitivity and the 1-specificity were 81.8% and 0%, respectively, for the identification of the CYP2C9 substrates in CYP2C9–CYP2D6 data set (Figure 5). A similar result was obtained in the identification of the CYP2D6 substrates in the same data set where the sensitivity and the 1-specificity were 84.4% and 4.7%, respectively (Figure 5). Thus, the selection of appropriate reference structures and the combination of their individual Tanimoto measures coming from different similarity search methods in a consensus score is a robust strategy to explore better the chemical space described by the set of reference

structures, and in this way increase the number of CYP substrates identified.

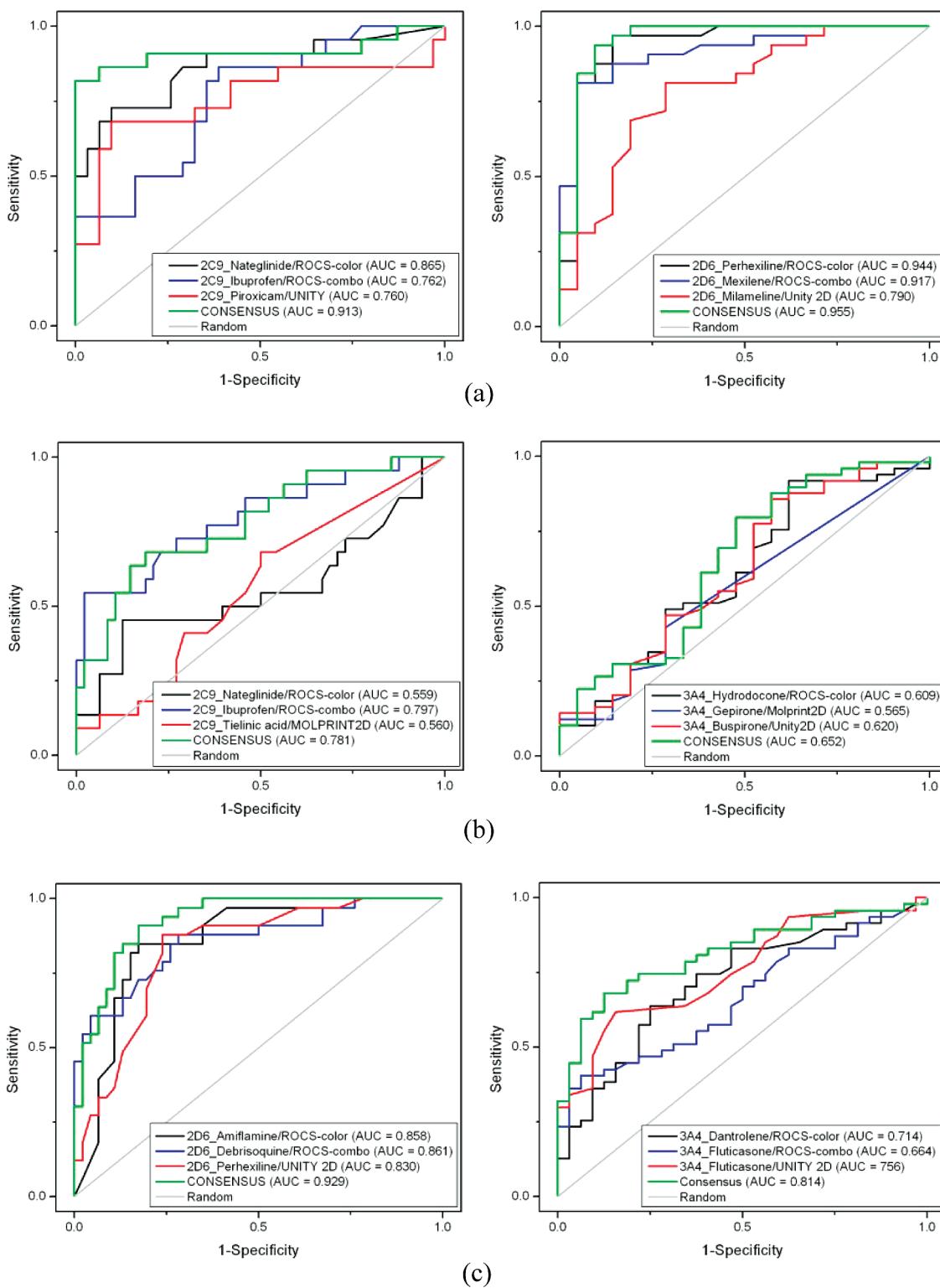
## CONCLUSION

In this work, we carried out pairwise similarity searches to assess the ability of 2D and 3D-similarity searches to identify reference structures with a high capacity to discriminate substrates and nonsubstrates for three pairs of data sets: CYP2C9–CYP2D6, CYP2C9–CYP3A4, and CYP2D6–CYP3A4. A detailed ROC-AUC analysis demonstrates that even for enzymes highly promiscuous like CYP450, the similarity search methods used in the present work were capable of identifying reference structures that are sufficiently representative of the whole substrates for a specific enzyme, with the best ones showing AUC values that range from 0.749 to 0.916.

The effectiveness of using single structures in similarity searching prompted us to use the combination of the Tanimoto scores of the best three reference structures into a new consensus Tanimoto score,  $T_{\text{consensus}}$ . The results of this analysis showed that by using multiple reference structures we achieved an improvement in both the area under the curve and the recall performance at the beginning of the list.

Our strategy was validated with an external data set of substrates using the best single reference structures and their combination. The performance, quantified by the AUC values, to discriminate between substrates and nonsubstrates was at the same level or superior to the results obtained during the development of the model.

Herein, we prefer not to establish threshold values for the Tanimoto coefficient to be used to predict the predominant isoform, because this might be the user's choice and resource. For instance, a conservative attitude (privileging specificity over sensitivity by requiring a high score) allows the majority



**Figure 5.** ROC curves for the validation data set: (a) CYP2C9–CYP2D6; (b) CYP2C9–CYP3A4; and (c) CYP2D6–CYP3A4.

of inactives (nonsubstrates) to be left aside. The liberal strategy (privileging sensitivity over specificity by requiring a high score) is a way to account for the uncertainty of models (see Triballeau et al.<sup>29</sup>).

Overall, we believe that the approach presented here, which uses different similarity search methods with multiple reference structures, will provide a powerful strategy in the identification of potential substrates for the CYP450 enzymes. We can also foresee its application on chemogenomics,

which entails the identification of compounds to be efficient in gene families.

## EXPERIMENTAL SECTION

**Data Set Preparation.** All compounds were subjected to a common preparation procedure, which included the following: addition of hydrogen atoms, determination of the most probable protonation state at pH 7.4, and generation

of up to three low energy conformers via the LigPrep modules of the Schrödinger software suite.<sup>35</sup>

Before running ROCS, the program OMEGA v.2.1 was used to convert all compounds to 3D multiconformer structures.<sup>36</sup> The algorithm implemented in OMEGA dissects the molecules into fragments, reassembles and regenerates many possible combinations, and then submits each conformer to a simplified energy evaluation. Next, all conformers below an energy threshold are compared, and those within a certain rms distance are clustered into one single representation. Default parameters were used with the following exceptions: (1) *buildff* (force field used for model construction and torsional search), this parameter was set as mmff94s (default = mmff94s\_NoEstat); (2) *maxconfs* (sets the maximum number of conformations to be generated), this parameter was set to 500 (default = 400); and (3) *rms* (sets the minimum root mean square (rms) distance below which two conformers are duplicates), this parameter was set to 0.6 (default = 0.8).

**ROCS Calculations.** A single low-energy conformation of each substrate in the database was used as the reference structure for ROCS.<sup>27,28</sup> In default operation, ROCS compares molecules purely on the basis of their best shape overlap, quantitated by their shape Tanimoto. It was quickly found that adding to the shape Tanimoto the score for the appropriate overlap of groups with comparable properties (donor, acceptor, hydrophobes, cation, anion, and ring), the so-called color score, and then ranking on this summed score improved virtual screening performance considerably. In this mode, ROCS optimizes the molecular overlay to maximize both the shape overlap and the color overlap obtained by aligning groups with the same properties that are contained in the color force field file. This overlay is then subsequently scored using the sum of shape Tanimoto for the overlay and the color score (the so-called combo score). The resulting database of pairwise CYP450 substrates, generated by OMEGA, was initially screened and scored using the ROCS algorithm to generate and score the 3D overlays of the database molecules. ROCS measures the shape similarity of two compounds by using the Tanimoto coefficient, which can vary from 0.0 to 1.0, with 1.0 representing an exact match.

**Molprint2D Calculations.** First, a hydrogen-depleted molecular structure for every substrate was constructed, and all heavy atoms had its Sybyl atom types assigned. Second, an all atom environment fingerprint was calculated, using distances from 0 to 2 bonds (from 0 to 3 bonds, in the case of the CYP2C9–CYP2D6 analysis), using the *mol22aejf.pl* pearl script. Next, the similarity of the compounds in the database was assessed to each reference structure using the *tanimoto.pl* pearl script.

**Unity 2D Calculations.** A sybyl database was constructed for each pairwise database. A single (optimized) conformation of each molecule, in the sybyl mol2 format, was put in the corresponding sybyl database. Next, a tripos molecular spreadsheet was created for each one of the three pairwise databases. We then used the sybyl script *tan\_columns* to create a Tanimoto distance array using Unity 2D fingerprints. The Unity 2D fingerprints column is created on the fly by this script.

**ROC-AUC Analysis.** In a database consisting of substrates of two CYP isoform, when a set was assigned as

substrate (set as 1 in the spreadsheet) of an enzyme, the other was identified as nonsubstrate (set as 0 in the spreadsheet) for the same enzyme and vice versa. For instance, to perform the ROC-AUC analysis in relation to the CYP2C9, the substrates of the CYP2D6 were assigned as nonsubstrate for the former enzyme. The opposite strategy was adopted when the analysis was with regard to the CYP2D6. The ROC graph and the calculation of the AUC values were performed using the ROC module of the SigmaPlot software.<sup>37</sup>

## ACKNOWLEDGMENT

We are grateful to the CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) and FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) for financial support and scholarships.

**Supporting Information Available:** Excel spreadsheet with AUC values of all cytochrome P450 substrates and the names of the substrates used in the external validation data set; Word document with all Tanimoto matrices for the selective pairwise comparisons and the ROC curves for the validation data set. This material is available free of charge via the Internet at <http://pubs.acs.org>.

## REFERENCES AND NOTES

- (1) Niggisch, F.; Mitchell, J. B. O. How to winnow actives from inactives: introducing molecular orthogonal sparse bigrams (MOSBs) and multiclass winnow. *J. Chem. Inf. Model.* **2008**, *48*, 306–318.
- (2) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods. *Drug Discovery Today* **2002**, *7*, 903–911.
- (3) Willert, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046–1053.
- (4) Kola, I.; Landis, J. Can the pharmaceutical industry reduce attrition rates? *J. Med. Chem.* **2004**, *3*, 711–716.
- (5) Tillement, J. P.; Tremblay, D. Clinical Pharmacokinetic Criteria for Drug Research. In *Comprehensive Medicinal Chemistry II: ADME-Tox Approaches*, 1st ed.; Triggle, D. J., Taylor, J. B., Eds.; Elsevier: Amsterdam, The Netherlands, 2006; Vol. 5, pp 11–30.
- (6) Wienkers, L. C.; Heath, T. G. Predicting *in vivo* drug interactions from *in vitro* drug discovery data. *Nat. Rev. Drug Discovery* **2005**, *4*, 825–833.
- (7) Lill, M. A.; Dobler, M.; Vedani, A. Prediction of small-molecule binding to cytochrome P450 3A4: Flexible docking combined with multidimensional QSAR. *ChemMedChem* **2006**, *1*, 73–81.
- (8) Iyer, R.; Zhang, D.; Zhang, D. Role of Drug Metabolism in Drug Development. In *Drug Metabolism in Drug Design and Development: Basic Concepts and Practice*, 1st ed.; Zhang, D., Zhu, M., Humphreys, W. G., Eds.; John Wiley & Sons, Inc.: Hoboken, NJ, 2008; Vol. 1, pp 261–285.
- (9) Zamora, I.; Afzelius, L.; Cruciani, G. Predicting Drug Metabolism: A site of metabolism prediction tool applied to the cytochrome P450 2C9. *J. Med. Chem.* **2003**, *46*, 2313–2324.
- (10) de Groot, M. J.; Ackland, M. J.; Horne, V. A.; Alex, A. A.; Jones, B. C. A novel approach to predicting P450 mediated drug metabolism. CYP2D6 catalyzed N-dealkylation reactions and qualitative metabolite predictions using a combined protein and pharmacophore model for CYP2D6. *J. Med. Chem.* **1999**, *42*, 4062–4070.
- (11) Bonn, B.; Masimirembwa, C. M.; Aristei, Y.; Zamora, I. The molecular basis of CYP2D6-mediated N-dealkylation: balance between metabolic clearance routes and enzyme inhibition. *Drug Metab. Dispos.* **2008**, *36*, 2199–2210.
- (12) Schuster, D.; Laggner, C.; Steindl, T. M.; Langer, T. Development and validation of an *in silico* P450 profiler based on pharmacophore models. *Curr. Drug Discovery Technol.* **2006**, *3*, 1–48.
- (13) Sheridan, R. P.; Korzekwa, K. R.; Torres, R. A.; Walker, M. J. Empirical regioselectivity models for human cytochromes P450 3A4, 2D6, and 2C9. *J. Med. Chem.* **2007**, *50*, 3173–3184.
- (14) Ekins, S.; Bravi, G.; Wikle, J. H.; Wrighton, S. A. Three-dimensional-quantitative structure activity relationship analysis of cytochrome P-450 3A4 substrates. *J. Pharmacol. Exp. Ther.* **1999**, *291*, 424–443.
- (15) Rendic, S. Summary of information on human CYP enzymes: human P450 metabolism data. *Drug Metab. Rev.* **2002**, *34*, 83–448.

- (16) Terfloth, L.; Bienfait, B.; Gastaiger, J. Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates. *J. Chem. Inf. Model.* **2007**, *47*, 1688–1701.
- (17) Yap, C. W.; Chen, Y. Z. Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *J. Chem. Inf. Model.* **2005**, *45*, 982–992.
- (18) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391–405.
- (19) Pavlidis, P.; Nobel, W. S. Matrix2png: a utility for visualizing matrix data. *Bioinformatics* **2003**, *19*, 295–296.
- (20) Hawkins, P. C. D.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to do an evaluation: pitfalls and traps. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 179–190.
- (21) Pham, T. A.; Jain, A. N. Parameter estimation for scoring protein-ligand interactions using negative training data. *J. Med. Chem.* **2006**, *49*, 5856–5868.
- (22) Sing, S. B.; Shen, L. Q.; Walker, M. J.; Sheridan, R. P. A model for predicting likely sites of CYP3A4-mediated metabolism on drug-like molecules. *J. Med. Chem.* **2003**, *46*, 1330–1336.
- (23) Kirchmair, J.; Ristic, S.; Eder, K.; Markt, P.; Wolber, G.; Laggner, C.; Langer, T. J. Fast and efficient in silico 3D screening: toward maximum computational efficiency of pharmacophore-based and shape-based approaches. *J. Chem. Inf. Model.* **2007**, *47*, 2182–2196.
- (24) Maréchal, J. D.; Kemp, C. A.; Roberts, G. C. K.; Paine, M. J. I.; Wolf, C. R.; Sutcliffe, M. J. Insights into drug metabolism by cytochromes P450 from modelling studies of CYP2D6-drug interactions. *Br. J. Pharmacol.* **2008**, *153*, S82–S89.
- (25) Groot, M. J.; Ackland, M. J.; Horne, V. A.; Alex, A. A.; Jones, B. C. A novel approach to predicting P450 mediated drug metabolism. CUP2d6 catalyzed n-dealkylation reactions and qualitative metabolite predictions using a combined protein and pharmacophore model for CYP2D6. *J. Med. Chem.* **1999**, *42*, 4062–4070.
- (26) Moffat, K.; Gillet, V. J.; Whittle, M.; Bravi, G.; Leach, A. L. A comparison of field-based similarity searching methods: CatShape, FBSS, and ROCS. *J. Chem. Inf. Model.* **2008**, *48*, 719–729.
- (27) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. L.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J. F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504–1519.
- (28) Kojej, T.; Engkvist, O.; Blomberg, N.; Muresan, S. Multifingerprint based similarity searches for targeted class compound selection. *J. Chem. Inf. Model.* **2006**, *46*, 1201–1213.
- (29) Hert, J.; Willett, P.; Wilton, D. J. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.
- (30) Hristozov, D. P.; Oprea, T. I.; Gasteiger, J. Virtual screening applications: a study of ligand-based methods and different structure representations in four different scenarios. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 617–640.
- (31) Nicholls, A. What do we know and when do we know it. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 239–255.
- (32) Truchon, J. F.; Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the “early recognition” problem. *J. Chem. Inf. Model.* **2007**, *47*, 488–508.
- (33) de Graaf, C.; Oostenbrink, C.; Keizers, P. H. J.; Wijst, T.; Jongejan, A.; Vermeulen, N. P. E. Catalytic site prediction and virtual screening of Cytochrome P450 2D6 substrates by consideration of water and rescoring in automated docking. *J. Med. Chem.* **2006**, *49*, 2417–2430.
- (34) Cytochrome P450 Drug Interaction Table. <http://medicine.iupui.edu/clinpharm/ddis/table.asp>; accessed June 10, 2009.
- (35) *LigPrep, version 2.0*; Schrödinger, LLC: New York, NY, 2005.
- (36) Böstrom, J.; Greenwood, J. R.; Gottfries, J. Assessing the performance of omega with respect to retrieving bioactive conformations. *J. Mol. Graphics Model.* **2003**, *21*, 449–462.
- (37) *SigmaPlot, version 10.0*; Systat Software Inc.: San Jose, CA, 2008.

CI900074T