

Reaction Site Mapping of Xenobiotic Biotransformations

Scott Boyer,^{*,†} Catrin Hasselgren Arnby,[†] Lars Carlsson,[†] James Smith,^{‡,§} Viktor Stein,^{‡,⊥} and Robert C. Glen[‡]

Safety Assessment, AstraZeneca Research & Development, 43183 Mölndal, Sweden, and Unilever Center for Molecular Sciences Informatics, University Chemical Laboratory, Cambridge, CB2 1EW, United Kingdom

Received August 28, 2006

Predictive metabolism methods can be used in drug discovery projects to enhance the understanding of structure–metabolism relationships. The present study uses data mining methods to exploit biotransformation data that have been recorded in the MDL Metabolite database. Reacting center fingerprints were derived from a comparison of substrates and their corresponding products listed in the database. This process yields two fingerprint databases: all atoms in all substrates and all reacting centers. The metabolic reaction data are then mined by submitting a new molecule and searching for fingerprint matches to every atom in the new molecule in both databases. An “occurrence ratio” is derived from the fingerprint matches between the submitted compound and the reacting center and substrate fingerprint databases. Normalization of the occurrence ratio within each submitted molecule enables the results of the search to be rank-ordered as a measure of the relative frequency of a reaction occurring at a specific site within the submitted molecule. Predictive performance that would allow this method to be used by drug discovery teams to generate useful hypotheses regarding structure metabolism relationships was observed.

INTRODUCTION

One of the common challenges in drug discovery is obtaining acceptable pharmacokinetics of candidate drugs. This process involves measuring and optimizing several parameters simultaneously to achieve the best balance of physicochemical properties and potency at the therapeutic target. One critical aspect of this optimization process is obtaining metabolic stability such that the candidate drug survives the passage over the gut wall and through the liver and reaches the target tissue(s).

Xenobiotic metabolism starts after absorption into the gut epithelia (presystemic metabolism), where metabolizing enzymes [for example, the heme–thiolate and monooxygenases—“cytochrome” (CYP) isoforms] are present in high concentrations. Once a xenobiotic has passed this barrier intact, it is exposed to an even more formidable defense system in the liver. Together, these drug metabolizing systems have evolved to effectively protect the body against exposure to excessive concentrations of “foreign” low-molecular-weight organic compounds present in ingested, dermally absorbed, and inhaled materials¹. In addition to the cytochrome P450s, however, there are a variety of other drug metabolizing enzymes that can affect how much of an orally administered drug reaches the systemic circulation including other oxidases, hydrolases, reductases, and dehydrogenases (oxidoreductases).

The metabolic modification of a molecule is described conventionally as a two-phase process, containing Phase I

functionalization reactions and Phase II biosynthetic reactions. Phase I reactions typically add polar functionality, resulting in either more polar metabolites or elimination reactions (e.g., dealkylation and deamination) often in preparation for Phase II, in which the metabolite is conjugated to a polar “carrier” that facilitates elimination through the kidneys or faeces. Typical conjugates are glucuronic acid, sulfate, glutathione, amino acids, or acetate. Conjugation reactions can also occur without prior Phase I functionalization reactions. Together, these systems, consisting of well over 50 different enzymes in humans, represent a large part of the xenobiotic defense system (the other large part of this system being xenobiotic transporters) and are generally characterized by three features (the three “D”s): *Diversity*, *Degeneracy*, and *Duplication*: *Diversity* in that a very diverse array of reactions can be carried out by this system, *Degeneracy* in that most members of this system can carry out reactions on a very wide variety of substrates, unlike “conventional” enzymes, and *Duplication* in that the same reaction can, at times, be carried out by several different enzymes.

Developing computational models of this very diverse system for use in problems of drug discovery has taken the form of using conventional molecular modeling techniques and/or developing rules to predict which metabolic reactions can take place on which chemical structures. This has been particularly the case for the CYP family of enzymes. A number of different predictive approaches have been reported and range from the development of quantitative structure–activity models or the deduction of pharmacophores to mixed quantum mechanics/structural studies in which a structural model of a P450 isoform was generated and used for docking and reaction modeling.^{2–11} Other approaches for exploring predictive metabolism have been rule-based using libraries

* Corresponding author tel.: +46 31 776 2882; e-mail: scott.boyer@astrazeneca.com.

[†] AstraZeneca Research & Development.

[‡] Unilever Center for Molecular Sciences Informatics.

[§] Current address: Computer-Chemie-Centrum, Institut für Organische Chemie, Universität Erlangen-Nürnberg, Nögelsbachstrasse 25, 91052 Erlangen, Germany.

[⊥] Current address: Department of Biochemistry, University of Cambridge, 80 Tennis Court Rd., Old Addenbrookes Site, Cambridge, CB2 1GA, U.K.

of molecular fragments involved in metabolic transformations.^{12–14} More sophisticated rule-based systems for metabolic transformation prediction based on biodegradation data sets have also been reported.^{15,16}

The present report is based on neither approximations of the metabolizing enzymes nor static reaction rules but rather uses reported reactions from the peer-reviewed literature to build a picture of the frequency of reported reactions at each atomic site of a new “query” compound. The method we have labeled the substrate product occurrence ratio calculator (SPORCalc) utilizes fingerprints of substrates reported in the Metabolite database of metabolic transformations, combined with reacting centers that are deduced from the biotransformation data.¹⁷ The fingerprints are hierarchical atom-type descriptions (also known as circular fingerprints) and have been implemented successfully in the prediction of pK_a values for organic acids and bases^{18,19} and more recently as measures of molecular similarity.^{20–22}

The objective of this study is to evaluate a method for facilitating the rapid mining of the data in Metabolite and the projection of those data onto a new structure, thereby developing a picture of the substructures most commonly reported to be involved in metabolic transformations. An assessment of the predictive power of this mining method is also presented.

METHODS

Metabolism Database. The Metabolite database version 2004.1 from MDL¹⁷ was used as a source of metabolic transformations of xenobiotics harvested from the literature. This version of the database contains 69 317 transformation files from published sources. Each file contains one or more individual transformations. Viewed by species, the largest fraction of transformations is 47% from rats followed by 32% from humans and 9% from dogs. The remaining fraction contains other mammalian species but does include a small proportion of nonmammalian species, including prokaryotes. The distribution of reactions processed and stored as fingerprints is 32 911 Phase I additions (hydroxylations and oxidations) and 71 456 eliminations (e.g., dealkylations and amide/ester hydrolysis reactions) and 26 817 Phase II additions (conjugations, etc.). Reaction types are given as one of the metadata fields in the database. The current method focuses on detecting all Phase I reactions for all species in one query, but it is easy to separate the data such that one search queries a certain species or reaction class only, for example, all rat hydroxylations or all human dealkylations.

All results are based on the following reaction classes: ester and amide hydrolysis reactions, aliphatic and aromatic hydroxylations, heteroatom oxidations, and any elimination reaction (ring openings, heteroatom dealkylations, and deaminations). In this study, all species are included in the data set.

Overview of the Method. The method is broken into two basic steps: (1) generation of the data sets, which includes (i) generation of the substrate data set and (ii) generation of the reaction center data set, and (2) the occurrence ratio calculator.

Generation of Data Sets. When using Metabolite, certain filters can be applied when exporting the data, for example,

species and reaction types, and hence different subsets of data can be investigated depending on specific interests. However, independent of previous filtering, any subset or complete set of exported data will be denoted “the entire data set”. The entire data set can be exported from Metabolite as an RD file. This file contains reaction pairs, where each substrate and product are listed in MOL format.¹⁷ Fingerprints for the substrate data set can be generated as described previously.^{18,19} These fingerprints are based on the SYBYL atom-type ordering and notation as generated by OpenBabel²³ and comprise 33 atom types.²⁴ Figure 1 gives a schematic overview of the fingerprint generation and an example of a six-level fingerprint. For the generation of reacting center fingerprints, however, the atoms involved in the reaction first need to be identified. This step is described below.

Metabolite contains annotations for reaction centers, although these are ambiguous in a large proportion of the data. Therefore, instead of using only the annotations, we compared the structures of the substrate and the product for every given transformation. Any reaction can be determined by looking at the union, that is, the maximum common substructure (MCS), of the substrate and the product, and deviations from the MCS in either the substrate or the product are usually sufficient requirements for the identification of a reaction center. It is possible to match those deviating substructures, which can occur in either the substrate or the product, with predefined substructures representing specific reaction classes, for example, dealkylation or oxygen addition. As an added benefit, this allows for a variable choice of reaction classes if one would wish to separate the data in this way.

To compute a MCS, we used the default settings for the MCS routine in OEChem²⁵ and compared it to the annotations in Metabolite. If the atoms in the MCS were consistent with the annotations, the MCS was accepted; otherwise, the settings in OEChem were changed to exact settings, resulting in a stricter matching. If the MCS still was not accepted, we used the MCS computed from OEChem with default settings without considering the Metabolite annotations at all. Once reaction centers are identified in this manner, fingerprints are generated for each atom in every substrate of the entire data set. The reaction centers in the substrates are identified and fingerprints generated which are stored separately as a “reaction center subset” of the entire data set, thus producing two separate databases: a database of substrate fingerprints and a much smaller database of reaction center fingerprints. In the present work, these databases are stored as ASCII files with the fingerprints listed sequentially.

Occurrence Ratio Calculator. New “query” compounds are converted to MOL format, and fingerprints are generated for each atom in exactly the same way that the fingerprints are generated for the substrates. The fingerprints in a new compound were then matched against the reaction center fingerprints and the entire data set fingerprints. By comparing the fingerprint matrices using combinations of two different similarity operators, an exact match operator and a distance operator, the number of hits in each data set is counted.

The exact match operator requires that corresponding rows in two fingerprint matrices are exactly the same, hence $F_j^a = F_j^b$, where j denotes the row. The distance between two

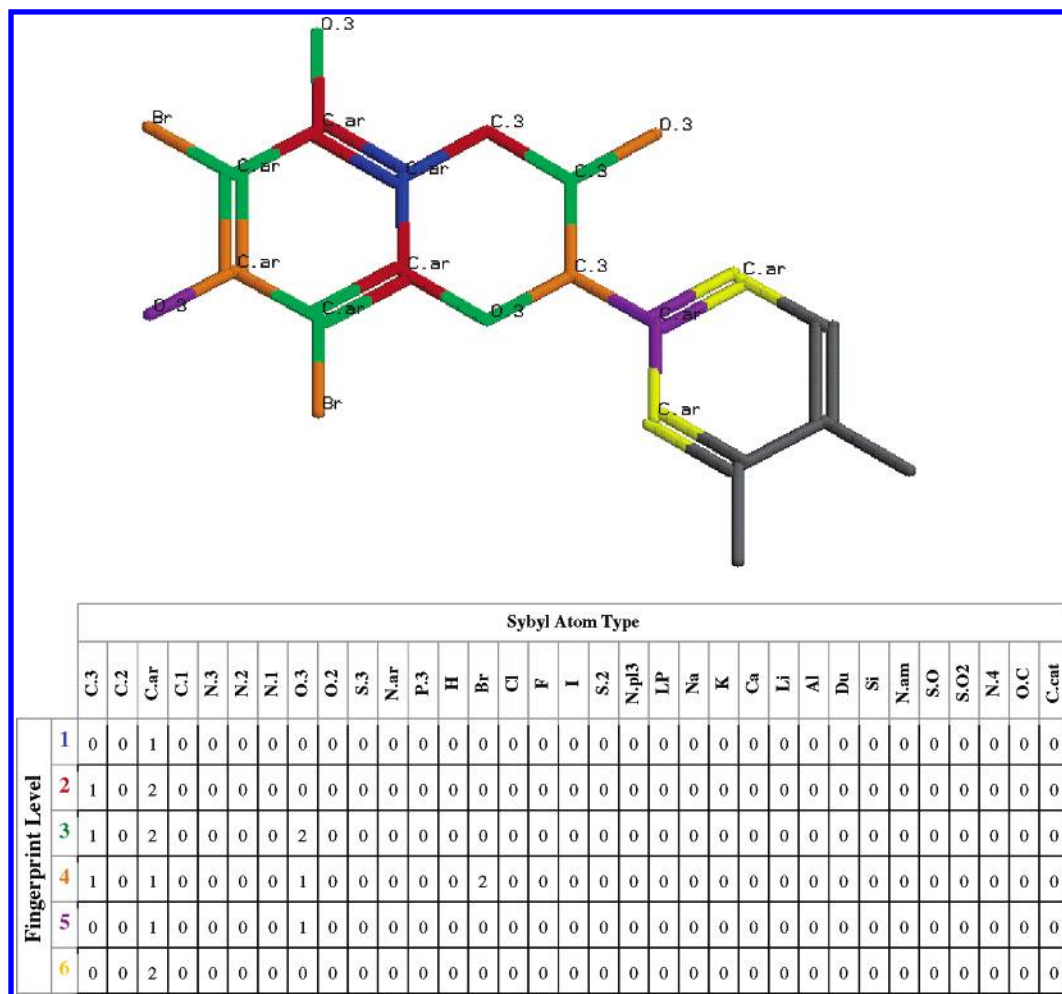


Figure 1. Illustration of a six-level fingerprint. The fingerprint is centered at the blue C.ar. The successive levels range from orange to yellow, green, blue, and cyan. Atoms that lie further away are not considered. Below: the corresponding fingerprint matrix where each column in the matrix represents a Sybyl atom type defined as C.3, C.2, C.ar, C.1, N.3, N.2, N.1, O.3, O.2, S.3, N.ar, P.3, H, Br, Cl, F, I, S.2, N.pl3, LP, Na, K, Ca, Li, Al, Du, Si, N.am, S.O, S.O2, N.4, O.CO2, or C.cat.²⁰ The rows are colored according to the same color scheme as that in the top part of the figure.

fingerprints a and b , for the j th row, is defined as

$$d_j = 1 - \frac{\sum_{n=1}^{33} F_{j,n}^a F_{j,n}^b}{\sum_{n=1}^{33} [(F_{j,n}^a)^2 + (F_{j,n}^b)^2 - F_{j,n}^a F_{j,n}^b]}$$

Using a distance as a similarity operator is only meaningful if a similarity threshold, s , is defined. Two fingerprints are considered to be equal if

$$s \geq \sum_{j=1}^6 w_j d_j$$

where w_j is a weight that can be used to adjust the significance of a row. Any combination of the two operators can be used, and we have utilized three combinations of these two operators to produce increasing grades of matching criteria for the fingerprints. The settings used in this study are labeled “Fuzzy”, “Standard”, and “Strict” to reflect the matching criteria, and the exact settings for the distance similarity and fingerprint row weighting are given in Table

Table 1. Fingerprint Matching Settings for SPORCalc used in the Present Study^a

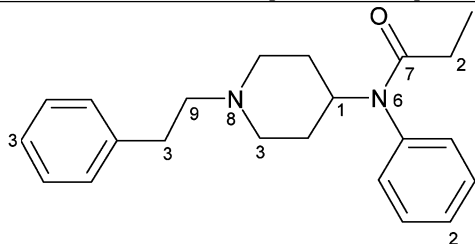
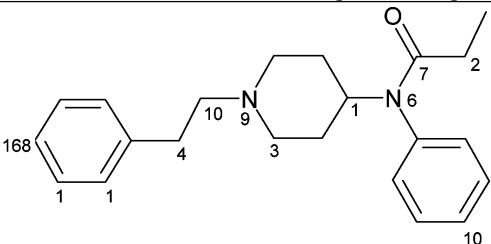
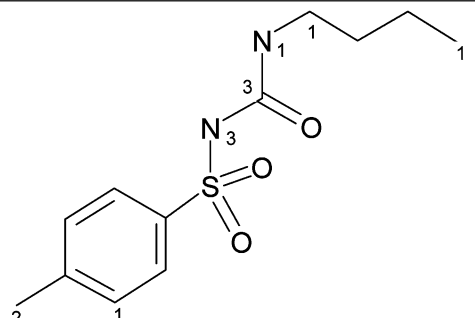
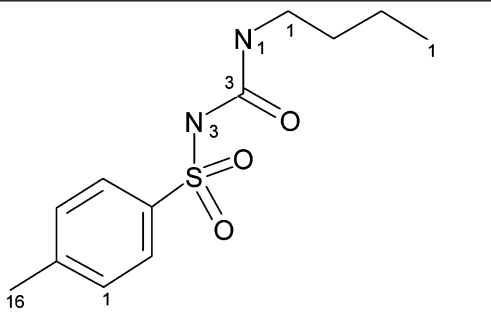
level/setting	Fuzzy similarity threshold, $s = 1.0$	Standard similarity threshold, $s = 0.5$	stRict similarity threshold, $s = 0.1$
1	exact	exact	exact
2	exact	exact	exact
3	1.0	exact	exact
4	0.75	0.75	exact
5	0.50	0.50	0.50
6	0.25	0.25	0.25

^a The fingerprint levels are specified with their respective matching or weightings. Generally, the fingerprint weighting decreases as the distance to the level 1 atom increases. Exact matches of fingerprints are required for the first two, three, or four levels for the Fuzzy, Standard, and Strict settings, respectively.

1. For computational efficiency, the operators are applied on one row at a time so that fingerprints that do not match in the upper levels can be rejected quickly without requiring further analysis.

With the fingerprint generation and the operators described above, the occurrence ratios, r_i , can be calculated and normalized. The normalized occurrence ratio, \bar{r}_i , for each

Table 2. Recovery of Reaction Center Information from the Metabolite Database^a

Compound	Number of reactions occurring when searching Metabolite*	Number of reactions determined using exact settings*
Fentanyl		
Tolbutamide		

^a The number of hits for a selected set of atoms using manual searches in the Metabolite database was recorded for relevant substructures in the two example compounds (left columns) and compared to the automated searching conducted in SPORCalc (right columns). Results indicate that no information was lost in transferring the metabolic reactions in Metabolite to fingerprints. Unlabeled atoms gave no hits. The manual search was performed using the “exact substrate structure” setting in Metabolite. Results are for two of the eight compounds assessed for verification of the method. For a list of all eight compounds, please see Methods.

atom in the new compound is calculated as described previously.⁹

$$\bar{r}_i = \frac{r_i}{\max_i (r_i)}$$

Visualization of the Normalized Occurrence Ratio. The normalized occurrence ratio is then projected as a color onto the structure of the new compound, in this case using RasMol.²⁶ The rules for color assignment can be varied to suit the requirements of the user, but in this report the following rules were applied: white, $0 \leq \bar{r}_i < 0.15$; green, $0.15 \leq \bar{r}_i < 0.33$; orange, $0.33 \leq \bar{r}_i < 0.66$; red, $0.66 \leq \bar{r}_i \leq 1.00$.

Implementation Details. The generation of data sets is not important in respect to execution times. This part of the method mainly deals with inputting and outputting the different file formats previously mentioned; hence, this part has been implemented using Python. The computation of occurrence ratios for a query molecule is critical to the user, since this is the time taken to get a result. This part of the algorithm has been implemented using C++ and compiled with the GNU compiler. Our test runs for a single-setting query took a maximum of 2 min on a computer with a 2.2 GHz AMD Opteron processor, with 2 GB of working memory and running Linux RHEL.

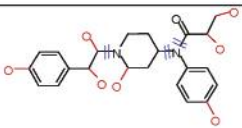
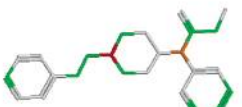

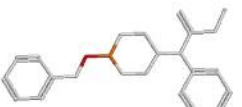
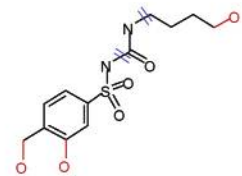



RESULTS

Verification of Database Content. Any database will contain a certain number of erroneous entries, and this could affect an occurrence frequency-based method such as that outlined in this study. Checking all entries in the Metabolite database for erroneous entries or ambiguous annotations is beyond the scope of this study, however. Instead, we adopted

a sampling procedure which is commonly used as a statistical tool in quality control. We have taken a random sample of 250 reactions and manually checked that the reaction center is correctly classified and that the correct atoms are involved in the reaction. No errors were found in this sample, which represents a confidence greater than 90% that there will be less than 1% misclassified reactions in the reaction data set.²⁷

Verification of Fingerprint Recognition. Eight compounds from the database were selected for confirmation of the method for determining fingerprints and counting hits (amitriptyline, almokalant, amodiaquine, tolbutamide, tamoxifen, amiodarone, fentanyl, and lansoprazole). The compounds were queried using exact settings during the fingerprint matching to directly compare the actual number of hits from a manual search in the original database with the number of hits found using the proposed method. In this implementation, the method considers atoms up to five bonds away from the reaction center atom, and so the manual search was tailored to find the matching fragment. For illustration, data from two of the eight tested compounds are presented in Table 2 and demonstrate that the algorithm for identifying reaction sites and counting the number of occurrences performs as intended. All reactions reported in Metabolite are correctly classified, and the atoms involved in the reactions from all eight compounds are marked, confirming that the method is reflecting the information content of Metabolite without losing data. It is possible to get a higher number of hits for an atom during a query, as there can be other substrates containing the same substructure, but the number of hits should never be lower than that obtained from a manual search in Metabolite. There are, for example, 168 hits on the carbon in the para position in fentanyl (Table 2) as it is a common environment for hydroxylation, and three of these hits are actually from fentanyl itself (which can be

Table 3. Color-Coded Results from Searching a Selected Set of Compounds with Known Metabolic Patterns (Column 1) Using the Fuzzy, Standard, and Strict Settings (Columns 2–4, Respectively) in SPORCalc^a

Compound	Experimental Sites	Fuzzy	Standard	Strict
Fentanyl				
Tolbutamide				

^a Red indicates the highest reported frequency in Metabolite and white the lowest (exact color-coding rules are given in Methods). Known hydroxylation sites (denoted by a red -O) and elimination sites (denoted by a blue hash through the bond) are those given for the respective compounds in Metabolite. The compounds have been queried against the database with all reactions involving the particular compound removed as leave-one-out searches. Results are for two of the eight compounds assessed for verification of the method. For a list of all eight compounds, please see Methods.

seen from the manual search of Metabolite). The purpose of this validation was to check that all reactions are picked up and correctly identified.

Once it was confirmed that the fingerprint matching procedure could accurately identify relevant substructures, the representation of occurrence ratios was investigated. Hence, for each compound queried, a subset of the Metabolite database in which all reactions involving that particular compound had been removed was used, in a “leave-one-out” fashion.

The metabolic sites used in this validation were confirmed by reference to the original literature and reflect the reported metabolic sites reported. The compounds were queried using three different similarity settings: Fuzzy, Standard, and Strict (Table 1). For illustration, the results from tolbutamide and fentanyl are presented in Table 3. The results of all eight compounds, when evaluated in this fashion, show that the metabolic sites identified correspond with experimentally determined sites and that the “Standard” setting gives a reasonably accurate result for the greatest number of sites.

“Prediction” of Metabolic Sites. In order to evaluate the predictive capacity of the method, 30 compounds not published in the 2004.1 version of Metabolite were selected and queried against fingerprint databases derived from this particular version. The 30 validation compounds were taken from the 2005.1 version of Metabolite. The objective with this operation was to verify how predictive the method is on unseen data. The compounds were run using standard settings, and the results are reported in Table 4.

To compare the prediction accuracy of SPORCalc to other published methods, we have utilized the system of quoting the predicted top three ranked sites, that is, the probability that the experimental site will be any of the predicted three most highly ranked sites.¹⁰ As our method indicates both additions and eliminations (detailed descriptions of “additions” and “eliminations” can be found in the Methods section), we had to establish certain criteria for ranking the sites. The occurrence ratio indicates the likelihood of reaction

for any atoms. For additions, the occurrence ratio can be directly used to compare the likelihood of metabolism at various atomic sites. For eliminations, however, both atoms involved are most often given a high occurrence ratio, and in these cases, both atoms are considered one site. In this case, the atom with the highest occurrence ratio is compared to the other sites. For example, if two atoms involved in an elimination are ranked 2 and 3, the site will be given rank 2 and the site ranked 4 will instead be ranked 3. The results of the 30 compounds compared show that, for Standard settings, 87% of the experimental sites are found within the three most highly ranked positions.

Interpretation. The reaction classes considered must be kept in mind when interpreting the results; a highlighted atom only indicates a reaction taking place; it says nothing about *which* reaction will take place. If two neighboring atoms are highlighted, for instance, it is quite likely that a bond breakage will occur, but it can also indicate that a bond breakage indeed occurs but that an oxidation also occurs (most often on aliphatic amines). This can be seen in amitriptyline, for example (Figure 2), where the nitrogen has a higher occurrence ratio indicating that it is involved in more reactions than the methyl groups. The sites are ranked against each other, and the normalized occurrence ratio provides a measure of the relative intracompound reaction site frequency. The user can therefore easily identify which sites are marked as the most probable.

During the evaluation of the “Fuzzy”, “Standard”, and “Strict” settings, it became apparent that the settings that one should use will depend on the query structure: the more common the environment, the stricter the settings one can use. In general, if one starts with the Standard settings, the subsequent runs can be made stricter or fuzzier depending on the number of hits found. A sensible way of using this method is to use the different settings and compare the results. A site that is marked as a probable site of reaction for all settings can be considered more reliable than one only appearing with Fuzzy settings. The actual number of hits in

Table 4. Color-Coded Results from Searching a Selected Set of Compounds with Known Metabolic Patterns (Column 1) Using the Standard Settings in SPORCalc^a

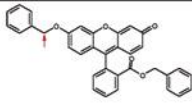
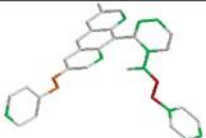
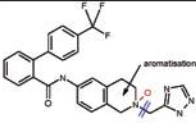
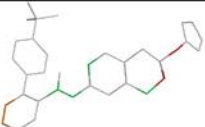
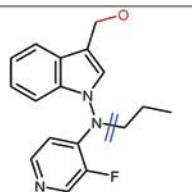
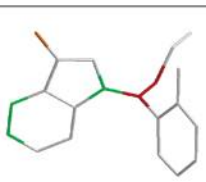
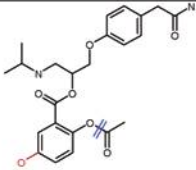
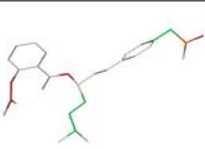
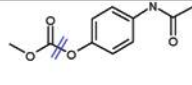
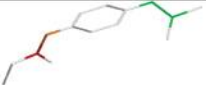
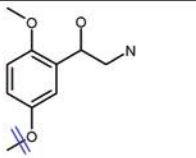
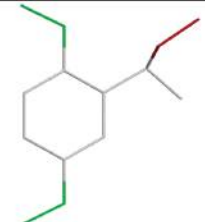
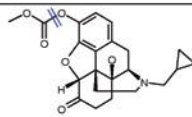
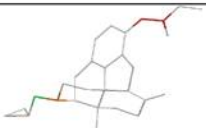
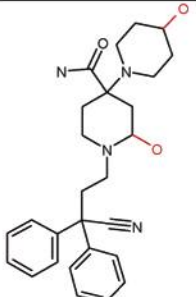
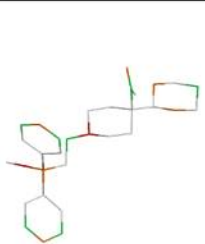
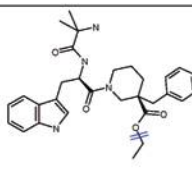
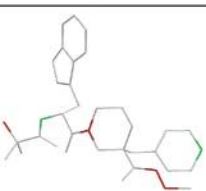
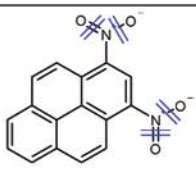
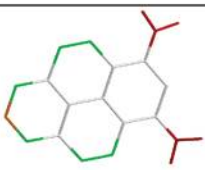
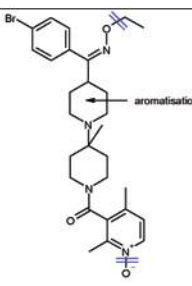
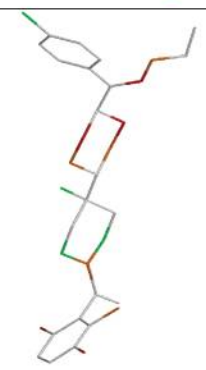
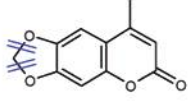
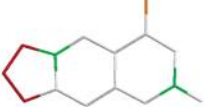
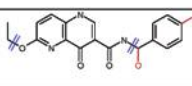
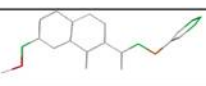
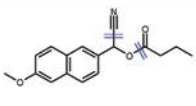
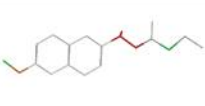
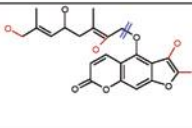
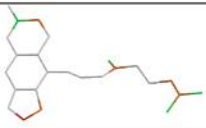
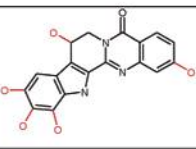
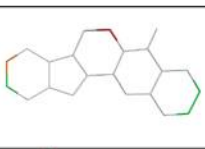
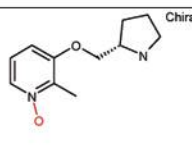
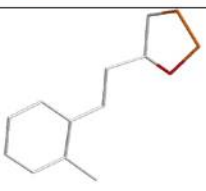
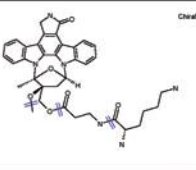
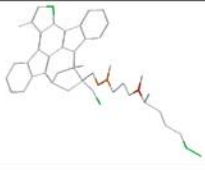
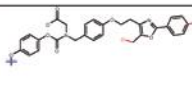
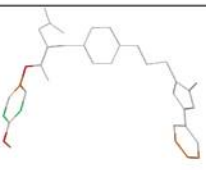
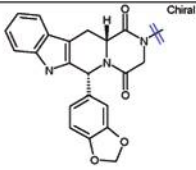
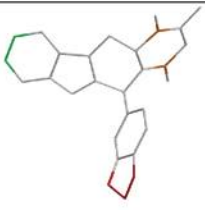
Dibenzylfluorescein (DBF) ²⁸			CP 346086		
HP 184			Atenolol aspirinate		
4-Methyloxycarbonyloxyacetanilide			Desglymidodrine		
Methyl naltrexone-3-O-carbonate			Piritramide		
1-(2(R)-(2-Amino-2-methylpropionylamino)-3-(1H-indol-3-yl)propionyl)-3-benzylpiperidine-3(S)-carboxylic acid ethyl ester			1,3-Dinitropyrene		
SCH 351125			6,7-Methylenedioxy-4-methylcoumarin		
CP 457920			(RS)-alpha-Cyano(6-methoxy-2-naphthyl)methyl butanoate		
Notopteron			Rutaecarpine		
ABT 089			CEP 2563		
Muraglitazar			Cialis		

Table 4 (Continued)

CT 51464			Sultamicillin		
4'-Bromoflavone			Depas		
4-((1-(Dicyclohexylacetyl)piperidine-4-ylidene)methyl)benzoic acid methyl ester			MK 0767		
Pyronaridine			6-Iodo-2-(4'-N-(3-fluoropropyl)methylamino)phenylimidazo(1,2-a)pyridine		
AM-630			Bergamottin		

^a Red indicates the highest reported frequency in Metabolite and white the lowest (exact color-coding rules are given in Methods). Known hydroxylation sites (denoted by a red -O) and elimination sites (denoted by a blue hash through the bond) are those given for the respective compounds in Metabolite. All 30 compounds are from a later version of Metabolite (except for DBF, which is not present in Metabolite at all) and are absent in the version used to query against.

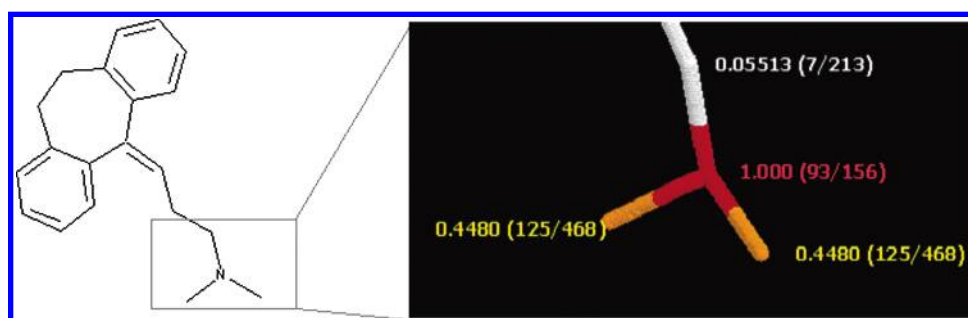


Figure 2. RasMol image of the reported normalized occurrence ratio and the number of “hits” from the reaction center database and the substrate database. This example is from the tertiary amine of amitriptyline and indicates that, for the amine nitrogen, the normalized occurrence ratio is 1.00 and the respective hits from the reaction center and substrate databases are 93 and 156, which constitutes the non-normalized occurrence ratio of 0.596.

the reaction center data can be displayed, as in Figure 2, and should also be examined as should the coloring. The coloring alone can be misleading in some instances in which very few examples of a relatively reactive center exist in the database.

DISCUSSION AND CONCLUSIONS

Here, we report automation of the previously reported manual counting of reactions in the Metabolite database to obtain relative reaction frequencies at atomic sites of a query molecule.⁹ The current method appears to be capable of

accurately extracting reaction-based information from a database like Metabolite, as well as reliably searching and displaying the results of a search for reaction centers. The proposed method has shown that it has the potential to utilize the complete information content in Metabolite and can be successful in identifying the putative reaction centers of compounds for which little or no metabolism information is available. It is important to note that the analysis is dependent on the diversity of reaction center environments, and thus a large database like Metabolite is useful in assuring a reasonable level of chemical diversity. It is critical that a database-mining method such as this should not lose information in the preprocessing steps, which would degrade the statistical advantage gained by searching thousands of reactions. The time gain for a user running this method compared to doing a manual search is quite significant.

An important benefit with this method is that it is independent of structural knowledge of the notoriously unspecific drug metabolizing enzymes. It is based solely on information related to how frequently a substructural environment is reported to be involved in a reaction. This is especially important early on in the drug discovery process when definitive information regarding the identity of the metabolizing enzyme is seldom available and can be used to compliment structure-based methods when information regarding the metabolizing enzyme is available. The method presented here could be extended to include chiral information on metabolic sites by inclusion of this information in the fingerprint generation step. This work is currently in progress. We have demonstrated that the method is able to correctly identify the three most probable sites of metabolism in 87% of the compounds—performance that is comparable to the best results of previously published methods. Real validation of sites-of-metabolism methods should be performed in the context of a drug discovery project and thus will vary depending on the application. This method, however, should be generally useful to drug discovery project teams trying to quickly establish structure—metabolism relationships in the absence of mass spectrometry or NMR data or in which available metabolite identification data are ambiguous.

ACKNOWLEDGMENT

J.S. and R.C.G. gratefully acknowledge the support of Unilever, AstraZeneca, and MDL for providing the Metabolite database.

REFERENCES AND NOTES

- (1) Parkinson, A. Biotransformation of Xenobiotics. In *Cassaret and Doull's Toxicology: The Basic Science of Poisons*; Klassen, C.D., Ed.; McGraw-Hill: New York, 1996; pp 113–186.
- (2) Korzekwa, K. R.; Jones, J. P.; Gillette, J. R. Theoretical Studies on Cytochrome P-450 Mediated Hydroxylation: A Predictive Model for Hydrogen Atom Abstractions. *J. Am. Chem. Soc.* **1990**, *112*, 7042–7046.
- (3) Mancy, A.; Broto, P.; Dijols, S.; Dansette, P. M.; Mansuy, D. The Substrate Binding Site of Human Liver Cytochrome P450 2C9: An Approach Using Designed Tienilic Acid Derivatives and Molecular Modelling. *Biochemistry* **1995**, *34*, 10365–10375.
- (4) Jones, B. C.; Hawksworth, G.; Horne, V. A.; Newlands, A.; Morsman, J.; Tute, M. S.; Smith, D. A. Putative Active Site Template Model for Cytochrome P450 2C9 (Tolbutamide Hydroxylase). *Drug Metab. Dispos.* **1996**, *24*, 260–266.
- (5) Korzekwa, K. R.; Grogan, J.; DeVito, S.; Jones, J. P. Electronic Models for Cytochrome P450 Oxidations. *Adv. Exp. Med. Biol.* **1996**, *38*, 361–369.
- (6) de Groot, M. J.; Ackland, M.; Horne, V.; Alexander, A.; Barry, J. Novel Approach to Predicting P450 Mediated Drug Metabolism. Development of Combined Protein and Pharmacophore Model for CYP2D6. *J. Med. Chem.* **1999**, *42*, 4062–4070.
- (7) Lewis, D. F.; Dickins, M.; Eddershaw, P. J.; Tarbit, M. H.; Goldfarb, P. S. Cytochrome P450 Substrate Specificities, Substrate Structural Templates and Enzyme Active Site Geometries. *Drug Metab. Drug Interact.* **1999**, *15*, 1–49.
- (8) Ekins, S.; Waller, C. L.; Swaan, P. W.; Cruciani, G.; Wrighton, S. A.; Wikel, J. H. Progress in Predicting Human ADME Parameters in Silico. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 251–272.
- (9) Boyer, S.; Zamora, I. New Methods in Predicting Metabolism. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 403–413.
- (10) 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* (12), 2313–2324.
- (11) Jones, J. P.; Mysinger, M.; Korzekwa, K. R. Computational Models for Cytochrome P450. A Predictive Electronic Model for Aromatic Oxidation and Hydrogen Atom Abstraction. *Drug Metab. Dispos.* **2002**, *30*, 7–12.
- (12) Talafous, J.; Sayre, L.; Mieyal, J.; Klopman, G. J. META. 2. A Dictionary Model of Mammalian Xenobiotic Metabolism. *J. Chem. Inf. Comput. Sci.* **2001**, *34*, 1326–1333.
- (13) Greene, N.; Judson, P. N.; Langowski, J. J.; Marchant, C. A. Knowledge-Based Expert Systems for Toxicity and Metabolism Prediction: DEREK, StAR and METEOR. *SAR QSAR Environ. Res.* **1999**, *10* (2–3), 299–314.
- (14) Darvas, F.; Marakhazi, S.; Kormos, P.; Kulkarni, G.; Kalasz, H.; Papp, A. Databases and High Throughput Testing during Drug Design and Development. In *Drug Metabolism*; Erhard, P. E., Ed.; Blackwell Science: Gamborg, U. K., 1999; pp 237–270.
- (15) Ellis, L. B. M.; Hou, B. K.; Kang, W.; Wackett, L. P. The University of Minnesota Biocatalysis/Biodegradation Database: Post-Genomic Data Mining. *Nucleic Acids Res.* **2003**, *31* (1), 262–265.
- (16) Mekenyan, O. G.; Dimitrov, S. D.; Pavlov, T. S.; Veith, G. D. A Systematic Approach to Simulating Metabolism in Computational Toxicology. I. The TIMES Heuristic Modelling Framework. *Curr. Pharm. Des.* **2004**, *10*, 1273–1293.
- (17) MDL ISIS Base; ISIS Draw; Metabolite Database; MDL Information Systems Inc.; San Ramon, CA. URL: <http://www.mdl.com/>. MOL file format is described in <http://www.mdl.com/downloads/public/ctfile/ctfile.pdf> (accessed Feb 2006).
- (18) Xing, L.; Glen, R. C. Novel Methods for the Prediction of pKa, logP and logD. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (4), 796–805.
- (19) Xing, L.; Glen, R. C.; Clark, R. D. Predicting pKa by Molecular Tree Structured Fingerprints and PLS. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (3), 870–879.
- (20) Bender, A.; Mussa, H. Y.; Glen, R. C. Molecular Similarity Searching using Atom Environments, Information-Based Feature Selection and a Naïve Bayesian Classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (1), 170–178.
- (21) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. J. Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1708–1718.
- (22) Bender, A.; Mussa, H. Y.; Gill, G. S.; Glen, R. C. Molecular Surface Point Environments Applied to Virtual Screening and the Elucidation of Binding Patterns. *J. Med. Chem.* **2004**, *47*, 6569–6583.
- (23) Open Babel. <http://openbabel.sourceforge.net> (accessed Feb 2006).
- (24) Clark, M.; Cramer, R. D., III; Van Opdenbosch, N. Validation of the General Purpose Tripos 5.2 Force Field. *J. Comput. Chem.* **1989**, *10*, 982–1012.
- (25) OpenEye Scientific Software – OEChem Toolkit. <http://www.eyesopen.com/products/toolkits/oechem.html> (accessed Feb 2006).
- (26) RasMol. <http://www.umass.edu/microbio/rasmol/> (accessed Feb 2006).
- (27) Montgomery, D. *Introduction to Statistical Quality Control*; John Wiley: New York, 1991.
- (28) de Groot, M. J.; Alex, A. A.; Jones, B. C. Development of a Combined Protein and Pharmacophore Model for Cytochrome P450 2C9. *J. Med. Chem.* **2002**, *45*, 1983–1993.

CI600376Q