DOI: 10.1002/cmdc.200700312

SyGMa: Combining Expert Knowledge and Empirical Scoring in the Prediction of Metabolites

Lars Ridder* and Markus Wagener*[a]

Predictions of potential metabolites based on chemical structure are becoming increasingly important in drug discovery to guide medicinal chemistry efforts that address metabolic issues and to support experimental metabolite screening and identification. Herein we present a novel rule-based method, SyGMa (Systematic Generation of potential Metabolites), to predict the potential metabolites of a given parent structure. A set of reaction rules covering a broad range of phase 1 and phase 2 metabolism has been derived from metabolic reactions reported in the Metabolite Database to occur in humans. An empirical probability score is assigned to each rule representing the fraction of correctly predicted metabolites in the training database. This score is used to refine the rules and to rank predicted metabolites. The current rule set of SyGMa covers approximately 70% of biotransformation reactions observed in humans. Evaluation of the rule-based predictions demonstrated a significant enrichment of true metabolites in the top of the ranking list: while in total, 68% of all observed metabolites in an independent test set were reproduced by SyGMa, a large part, 30% of the observed metabolites, were identified among the top three predictions. From a subset of cytochrome P450 specific metabolites, 84% were reproduced overall, with 66% in the top three predicted phase 1 metabolites. A similarity analysis of the reactions present in the database was performed to obtain an overview of the metabolic reactions predicted by SyGMa and to support ongoing efforts to extend the rules. Specific examples demonstrate the use of SyGMa in experimental metabolite identification and the application of SyGMa to suggest chemical modifications that improve the metabolic stability of compounds.

Introduction

Identification of metabolites is an important aspect of drug discovery and development at various stages in the process.[1] Early in lead discovery, metabolite identification is often required to support the chemical optimization toward metabolically stable compounds. Later in drug development it is essential to investigate the metabolic profile of a compound and to study the potential activity or toxicity of major metabolites. Predictions of metabolites can assist these activities in several ways. Early metabolite screening can be facilitated significantly by predictions. For example, fast LC–MS scans can be carried out to specifically detect predicted metabolites,[2] allowing relatively simple experimental setup and data analysis. Prediction methods can be helpful subsequently when interpreting the results and assessing possible chemical modifications to block metabolically labile sites. Furthermore, recent developments demonstrate that metabolite prediction in combination with MS data prediction can be used to facilitate the analysis of complex LC–MS–MS data resulting from full metabolite identification experiments.[3]

Many different methodologies to predict metabolites or sites of metabolism have been reported recently. The metabolic fate of a molecule depends on its chemical reactivity toward several metabolic processes that can occur, as well as on its interactions (affinity and binding orientation) with the biotransformation enzymes involved. An important approach in the prediction of metabolites is based on explicit calculation of (relative) chemical reactivities of different sites in a molecule and/or prediction of the binding of the molecule to metabolic enzymes. It is well established that calculated energies of hydrogen radical abstraction (e.g. by approximate quantum chemical methods) are a useful indicator of the metabolic lability of various aliphatic positions toward a range of cytochrome P450 catalyzed reactions.[4–6] Other calculations are used to assess the regioselectivity of aromatic hydroxylations by P450 enzymes.[4] Frontier orbital theory or Fukui calculations have been applied to predict the regioselectivity of aromatic hydroxylation[7] or to identify metabolically labile sites in complete molecules.[8] Docking has been used to predict the binding mode of ligands for CYP 2D6,[9] and the predicted exposure to the reactive heme cofactor was shown to correlate with the known sites of metabolism of the ligands. In a less explicit approach, a GRID-based (binding) interaction pattern of the CYP 2C9 active site was matched with those of its substrates to predict likely sites of metabolism.[10] The program MetaSite[11] combines the latter method to account for binding to several P450 isoenzymes with an approximate QM-based estimate of the reactivity of the individual atoms in a molecule to predict the site of metabolism. The above "first principles" approaches to predict metabolism are based on the chemical

[a] *Dr. L. Ridder, Dr. M. Wagener*
Molecular Design and Informatics
Organon, part of Schering-Plough Corporation
P.O. Box 20, 5340 BH Oss (The Netherlands)
Fax: (+31) 412 662539
E-mail: lars.ridder@organon.com
markus.wagener@organon.com

structures of ligands and catalytic sites, and not on prior experience or training sets. Therefore, they can potentially make useful predictions for new compound classes with unknown metabolic profiles. Most of these approaches are, however, limited to P450 catalyzed reactions and only indicate labile sites, rather than predicting the actual metabolites formed. Consequently, these methods are less suitable for routine use to support experimental metabolite identification.

Rule-based methods rely on metabolic rules derived by experts. Examples are Meta,[12,13] MetabolExpert,[14] Meteor,[15,16] Metadrug,[17–19] and KnowItAll.[20] These methods have the advantages of being potentially fast and generating actual structures of metabolites. However, because large sets of metabolic rules are being applied, these methods often generate large numbers of unobserved metabolites, which limits their value to chemists in identifying labile sites in a molecule. Some of the above methods have implemented a differentiation between likely and unlikely metabolites, for example, by using prioritization of rules[21] or a reasoning model.[15]

A recent development is to apply statistical analysis on a large database of experimental metabolic reactions. Based on such analysis, empirical probabilities are obtained which indicate the likeliness of certain metabolism. Several methods have been reported that use empirical scoring of metabolism outcomes. The PASS-BioTransfo program provides a likeliness that a certain class of biotransformation reaction will occur.[22] The SPORCalc approach ranks sites in a molecule according to likeliness of undergoing metabolism.[23] Also, a number of methods have been described, such as TIMES[24] and Metadrug,[19] that provide a probability of predicted metabolites to be formed.[25]

We report a new approach based on reaction rules that are statistically evaluated on the basis of a large dataset of experimental data. Based on this analysis, empirical probability scores are calculated. The most important distinction from existing approaches as described above is that the rules are also modified and optimized on the basis of these probability scores in order to decrease the number of incorrect predictions and to distinguish between more and less metabolically labile groups. The resulting prediction tool, which we call SyGMa (Systematic Generation of potential Metabolites), ranks predicted metabolites based on the empirical probability scores. It combines the advantages of systematically generating actual metabolite structures, at low computational cost, with a good differentiation between more and less likely metabolic routes.

Results

Development of the rule base

A training dataset of 6187 fully characterized metabolic reactions observed in human studies was retrieved from the MDL Metabolite Database. This dataset was used to derive and manually optimize an extensive set of metabolic rules. Initial rules were based on common knowledge of metabolic reactions or visual inspection of frequently occurring reactions in the experimental dataset, or they were obtained from a systematic analysis of the dataset using reaction fingerprints (see below). Each individual rule was applied to all reactants in the dataset. The fraction of the resulting metabolites that match actual metabolites observed in humans was taken as an empirical probability score assigned to the rule. Upon inspection of the matching metabolic reactions, a rule was further refined to increase the probability ratio, or split into multiple rules to account for variations in reactivity of different reaction centers toward the specific reactions (see Experimental Section for more details). As an example, Figure 1 illustrates the division of

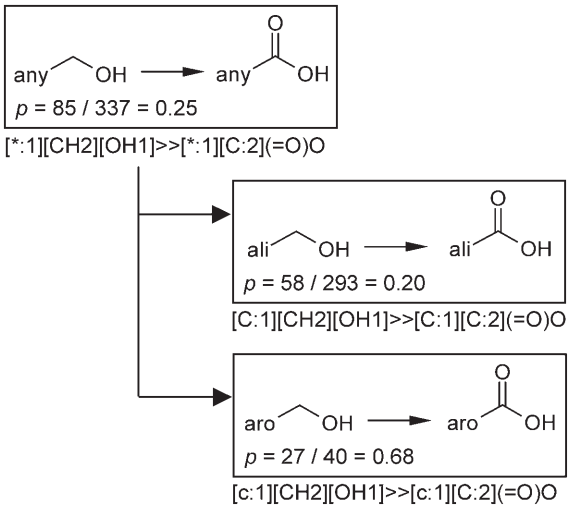

Figure 1. Division of a general rule for oxidation of a primary alcohol into two more specific rules, one for oxidation of an aliphatic primary alcohol and one for oxidation of a benzylic primary alcohol. Corresponding SMIRKS codes as well as calculated probability scores are shown.

a general rule for the oxidation of a primary alcohol into two more specific rules. One rule for oxidation of an aliphatic primary alcohol was created, which matches 58 of the initial 85 experimental examples of primary alcohol oxidation in the training set. The second rule for oxidation of a benzylic primary alcohol covers a smaller number of experimental examples, however, with a significantly higher probability score than the rule for primary alcohol oxidation. The splitting of the initial rule clearly resulted in new rules that account for the higher susceptibility of benzylic alcohols toward oxidation relative to aliphatic alcohols.

Another example of rule refinement is the division of an initial rule of O-glucuronidation of primary oxygen atoms into four more specific rules. These rules account for the observations that carboxyl oxygens are glucuronidated more frequently than hydroxy oxygens, and that both groups appeared to be more susceptible to glucuronidation when attached to aromatic cores than when they are attached to aliphatic groups. These differences in chemical environment will influence the nucleophilicity and acidity of the reacting oxygen centers. The effects on the observed frequencies can be understood given that glucuronidation proceeds via nucleophilic attack of the oxygen on UDP-glucuronic acid, and that the oxygen is activated through deprotonation by an active site base.[26] In similar

ways, distinctions could be made between more and less reactive chemical subgroups for most of the various types of metabolic reactions covered in the SyGMa rules.

The current rule base[27] contains 144 rules covering both phase 1 and 2 metabolism, with calculated probability ratios varying from 0.009 to 0.85. Figure 2a shows the distribution of the rules for the various types of metabolic reactions. The largest groups of rules are for dealkylation, carbon hydroxylation, and other types of carbon oxidation. These groups involved most extensive refinement during rule development, and some key aspects of this refinement are summarized hereafter.

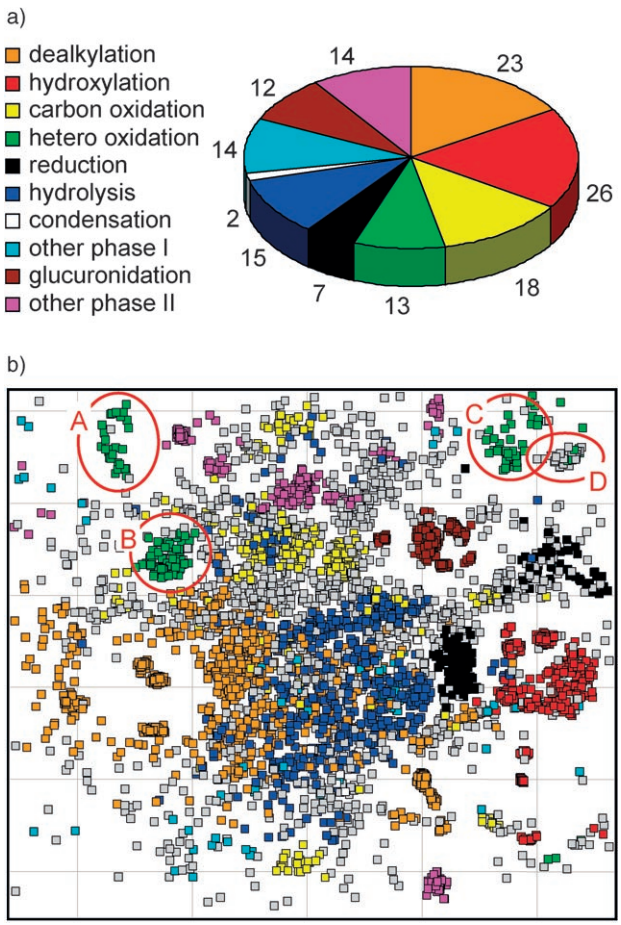

Figure 2. a) Distribution of the rules for the various types of metabolic reactions. b) Projection of all reactions in the training set on a 2D plane to optimally reflect reaction fingerprint distances calculated between all pairs of reactions. For each reaction, it was verified whether SyGMa could reproduce the metabolite in up to three subsequent reaction steps at a certain point during rule development. Reactions reproduced by SyGMa in two or three steps were excluded from the analysis. Reactions reproduced by SyGMa in one step are colored according to the rule they matched, within one of the indicated categories. Reactions not matched by SyGMa in up to three subsequent reaction steps are represented by gray dots. Four clusters of heteroatom oxidation reactions (A–D) are circled and exemplified in Figure 3.

The largest group of rules, that of hydroxylation reactions, comprises rules for aromatic (10), aliphatic (12), and benzylic hydroxylation (4). Hydroxylation reactions could be naively applied to any aromatic or aliphatic carbon atom in a molecule that is not fully substituted, potentially leading to many incorrectly predicted metabolites. Therefore, it was particularly important for this category of reactions to refine and split rules to decrease the number of predictions and to distinguish between more and less likely hydroxylation products. In the rules for aromatic hydroxylation, distinctions were made, for example, between positions para, meta, or ortho to other substituents and whether these substituents are connected through a carbon, oxygen, nitrogen, or other non-hydrogen atom. In addition, a priority of application was encoded in the order of para, ortho, and meta hydroxylation. In the rules for aliphatic hydroxylation, distinctions were made, for example, between primary, secondary, or tertiary aliphatic carbon atoms and whether these aliphatic carbon atoms were connected to other aromatic, conjugated or primary, secondary, tertiary or quaternary aliphatic carbon atoms, or to heteroatoms, etc. An additional distinguishing feature applied in the different rules is whether aliphatic carbon atoms are part of a ring or not.

The group of dealkylation rules includes relatively many different rules for N-dealkylation (i.e. 16), with probabilities ranging from 0.04 (N-dealkylation of piperazine) to 0.83 (N-demethylation of methylamine attached to an aromatic carbon atom). The probabilities within this group show internal consistency in that amino groups connected to aromatic carbon atoms are always more likely to dealkylate than amino groups attached to aliphatic groups only.

Carbon oxidation includes a number of rules for the formation of carboxyl groups. Carboxylation of primary carbon atoms can result from hydroxylation and subsequent oxidation. The individual steps are encoded in the rule set; however, the probabilities calculated for the metabolites resulting from these steps (by multiplication of the two individual probabilities as explained below) were relatively low. Because the two-step oxidation of primary carbon atoms to carboxylic acids is often detected and represented in the training dataset as single metabolic reactions, one-step rules for carboxylation of primary carbon atoms were included. These rules show significantly higher probabilities than would be obtained from applying the individual hydroxylation and oxidation steps. Note: as both the individual steps and the combined rules are included in the rule set, the carboxyl groups can be formed via two different pathways. The present method resolves this redundancy by selecting the path corresponding to the highest probability, which automatically results in the selection of the most appropriate rules.

The results of a "reaction similarity" analysis of the complete dataset is presented in Figure 2b. In this graph, each dot represents a reaction of the training dataset, such that the distance between each pair of dots approximately reflects the similarity of the corresponding reactions. The similarity is evaluated on the basis of reaction fingerprints as detailed in the Experimental Section. Colored dots are experimental reactions that are reproduced by SyGMa rules in one of the indicated categories at a certain point during rule development. Gray dots represent metabolic reactions that are not reproduced by SyGMa, in up to three subsequent reaction steps, with the existing rule base.

In most cases, reactions in the database that match a single rule have similar reaction fingerprints, and thus cluster together in Figure 2b. Some reaction types, such as hydrolysis and N-dealkylations, span larger differences in reaction fingerprints, as they involve removal of highly dissimilar parts of the molecules. Many reaction types, however, such as deacetylations and demethylations appear as clear clusters in the analysis presented in Figure 2b. Dense clusters of gray dots are a strong indication of a group of metabolic reactions that could be described with a new rule. To illustrate this, four clusters are highlighted in Figure 2b, three matching existing rules and one not matching an existing rule. Examples of reactions in each cluster are shown in Figure 3. Cluster D was found to consist of sulfide

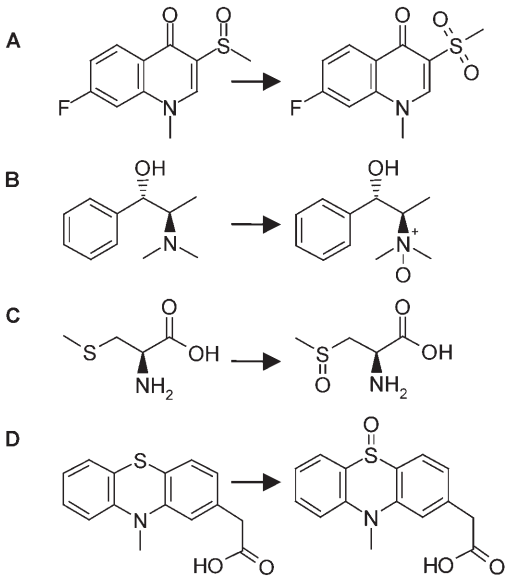

Figure 3. Typical examples of four clusters of reactions highlighted in Figure 2b. Cluster D identified sulfide oxidation reactions not yet covered by the existing rules of sulfide oxidation. Based on this finding the rule base was extended.

oxidation reactions, which were not covered by the existing rules for sulfide and sulfoxide oxidation matching clusters A and C. Based on this finding the rules could be extended to cover cluster D as well. This example shows the reaction fingerprint analysis to be a useful tool for building up the rule base. The remaining gray dots in Figure 2b are rather scattered, in agreement with the fact that most of these reactions are unique cases. Building rules for such reactions often leads to rules with low probability scores, resulting in mostly false predictions, or to rules lacking sufficient examples to derive meaningful probability scores. Also, part of the gray dots represent reactions that could have been reproduced by SyGMa in more than three subsequent reaction steps.

By themselves, the probabilities calculated for the individual rules provide useful information to chemists looking for modifications to improve metabolic stability within a chemical series. To illustrate the information contained in the rules and their corresponding probabilities, Figure 4 presents the top 10 "most probable" reactions of the current rule set. These rules are most likely to generate true biotransformation metabolites when they apply to a chemical structure. It is remarkable that the rules in this top 10 represent modifications of well-defined small functional groups. They provide a practical list of chemical features to avoid in a search for metabolically stable compounds. On the other hand, these most probable reactions can give useful suggestions for potential prodrugs that may be selectively metabolized in vivo into an active compound.

Evaluation of the rule-based predictions

SyGMa predicts metabolites by systematically applying all metabolic rules in the rule set described above, for a specified number of subsequent reaction steps. The metabolites are assigned the probability of the rule it was formed by, or the product of probabilities in case of multistep metabolites. Finally, the metabolites are rank-ordered by probability. When multiple subsequent reaction steps are applied, SyGMa, like other rule-based methods, potentially produces large numbers of metabolites. Therefore, the quality of the predictions needs to be measured not only in terms of the ability to reproduce experimental metabolites, but also in terms of enrichment of these true metabolites in the top of the ranking list. The performance of SyGMa was evaluated on the training set, as well as on an independent test set originating from a recent update of the MDL Metabolite Database.

A general difficulty in the evaluation of metabolite predictions is the considerable variability in the data for different

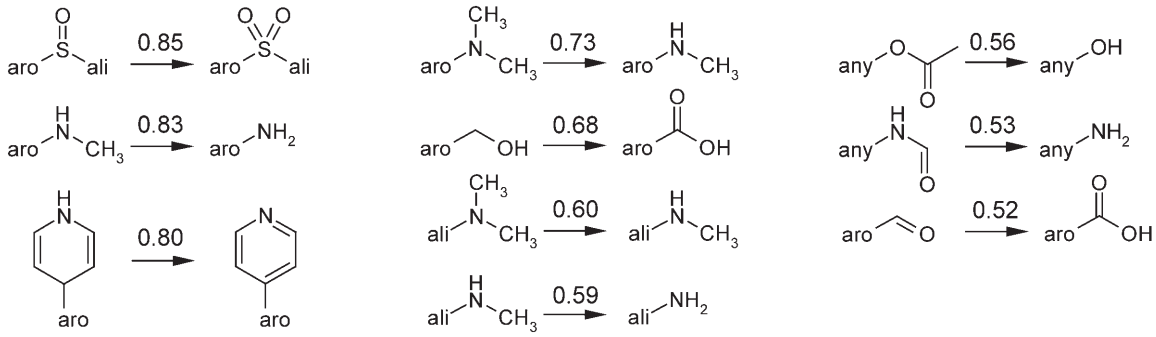

Figure 4. Top 10 "most probable" reactions; calculated probability factors are indicated above the arrows.

parent molecules. Some compounds have been extensively studied, resulting in the presence of numerous metabolites in the datasets (up to 33 metabolites for one parent compound in the training set). As a result, even in the case of a perfect prediction, up to rank 33 would have to be considered to reproduce 100% of the metabolites. For many other compounds, on the other hand, only a single metabolite is reported in the datasets. Figure 5 presents a distribution of metabolites per parent in the two human in vivo data sets. There is a clear bias toward compounds with a single metabolite, and it seems likely that for many of these compounds the reported metabolite profiles are not complete. This effect appears to be more pronounced in the test set in comparison with the training set. On average, 3.4 metabolites are reported per parent compound in the training set, and only 2.2 metabolites per parent compound in the test set. The compounds in the test set are more recent and probably less extensively studied than the compounds in the training set. Such imbalance can bias the evaluation results and make the interpretation more difficult.

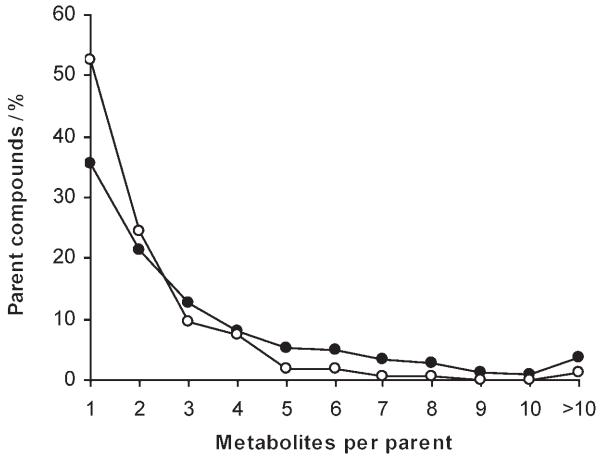

Figure 5. Distribution of metabolites per parent compound in the training (●) and test (○) sets of human in vivo metabolite data.

When SyGMa was applied to all parent compounds in the training set, 71% of the metabolites in the data set were reproduced. The fraction of major metabolites (metabolites in the database that are annotated "Major" on the basis of at least one referenced publication, suggesting these are quantitatively the most important metabolites) that is reproduced is even higher: 77%. These matches come from a large number of predicted metabolites generated by systematically applying all 144 rules to the parent compounds for up to three subsequent reaction steps. Figure 6a indicates the percentage of all experimental metabolites in the training set that are reproduced (i.e. the recall) as a function of the number of metabolites from the top of the ranking list that are taken into account: 44% of all experimental metabolites are reproduced within the top 10 predicted metabolites (——); this includes 54% of the major metabolites (– – –). The dotted line (-----) shows the percentage of the predictions that are experimentally confirmed (i.e. the precision): 37% of the predicted metabolites at rank 1 are experimentally observed and 28% of the top three ranked metabolites. When interpreting this number it should be taken into account that for the majority of the compounds, fewer than three metabolites have been reported and therefore, even with an ideal ranking method, the precision of the top three predictions could not be 100%.

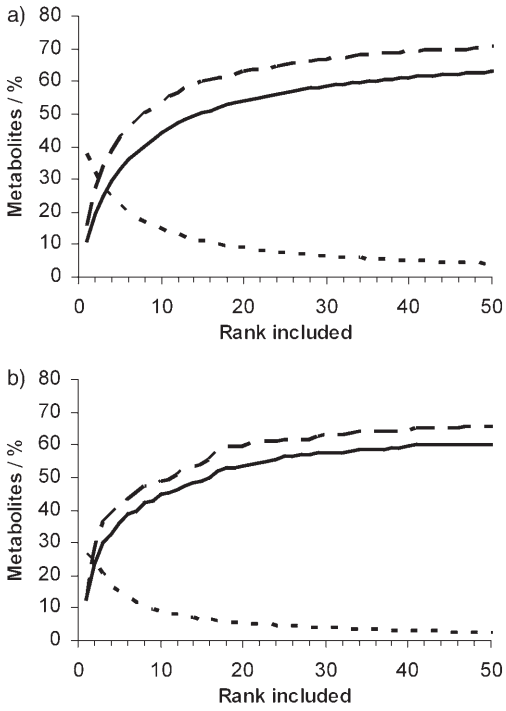

Figure 6. Percentage of experimental metabolites that are reproduced (recall; —— and – – –) and the percentage of the predictions that are experimentally confirmed (precision; -----) in a) the training set and b) the test set as function of the number of metabolites from the top of the ranking list that are taken into account. Solid lines represent the results for all metabolites, dashed lines (– – –) for major metabolites only.

The performance on the test set is very similar to the performance on the training set: 68% of the metabolites (69% of the major metabolites) are reproduced, and 45% of the metabolites are ranked in the top 10 (Figure 6b) including 49% of the major metabolites. The similarity in performance on the training data and test data indicates the robustness of the prediction method and the rule base. The precision is somewhat lower for the test set. This can be explained at least partly by the fact that the test set contains fewer metabolites per parent compound than the training set as shown in Figure 5.

To illustrate the diversity and complexity of the metabolic pathways used in the evaluation, Figure 7 illustrates the most important metabolic reactions of an extensively studied[28] example from the test set, lumiracoxib. The main metabolic modifications involve aromatic hydroxylation of the dihalophenyl ring (para to amine), oxidation of the benzylic methyl group to primary alcohol and carboxylic acid, glucuronidation of the carboxyl group, and condensation of the amine and carboxyl groups resulting in ring closure. These modifications were all predicted in the top five of the ranking list. Many multistep metabolites, involving combinations of these modifica-

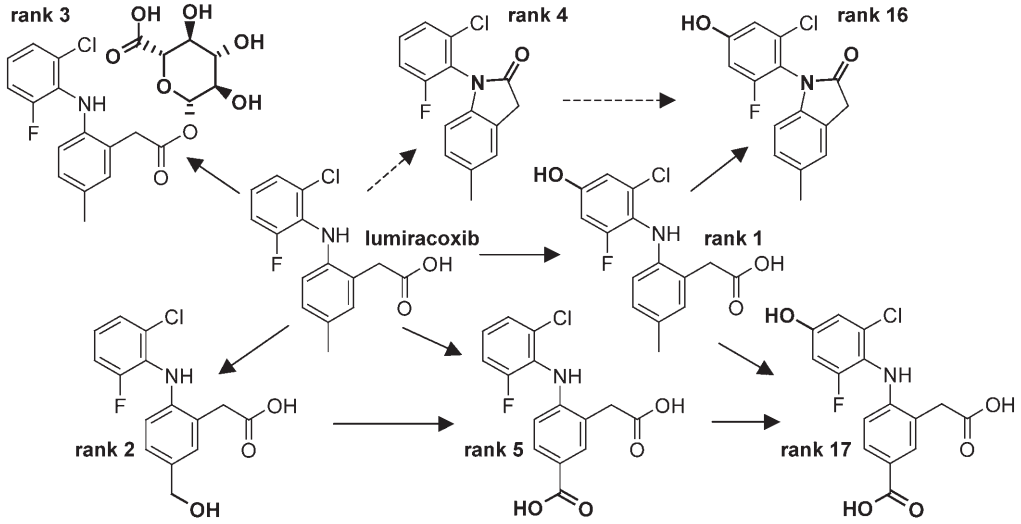

Figure 7. Summary of the most important metabolic reactions of lumiracoxib, a thoroughly studied example from the test set of metabolic reactions.[28] Ranks at which the metabolites were predicted by SyGMa are indicated. The reaction center is indicated in boldface. Note: ring closure, as predicted by SyGMa at rank 4 (indicated by dashed arrows) is not reported for the parent directly. For several metabolites, however, such as hydroxyphenyl (predicted at rank 1), the ring closure product (predicted at rank 16) is reported.

tions, are also reported. However, because probabilities of these metabolites are calculated as the product of the probabilities of the individual steps, they end up further down the ranking list of predicted metabolites. In total, all 15 fully characterized metabolites of lumiracoxib reported in the database are reproduced by SyGMa.

An important application of metabolite prediction in pharmaceutical research is to provide information on the basis of which useful suggestions for chemical modifications can be made to improve the metabolic profile in a lead series. Especially cytochromes P450 (CYPs) are generally considered as the most important family of metabolic enzymes that determine the metabolic stability of druglike molecules. To evaluate the usefulness of SyGMa specifically for predicting P450 metabolism, 127 single-step P450 reactions were selected from the test set, that is, reactions indicated in the Metabolite Database to be metabolized by one or more CYP isoenzymes. A subset of 118 SyGMa rules, covering only phase 1 metabolism, was applied in a single step to the 106 reactants of this dataset. Figure 8 shows the fraction of the 127 metabolites that are reproduced as a function of the ranks considered in the prediction; 107 experimental metabolites, that is, 84%, were reproduced by SyGMa, and 66% of the metabolites were predicted in the top three of the ranking list. This indicates that SyGMa, despite the fact that it has not been trained specifically to reproduce CYP metabolites, and without taking into account which CYP isoenzymes are involved in the metabolism, is capable of identifying most of the relevant CYP-catalyzed metabolic reactions in the top three of the ranking list.

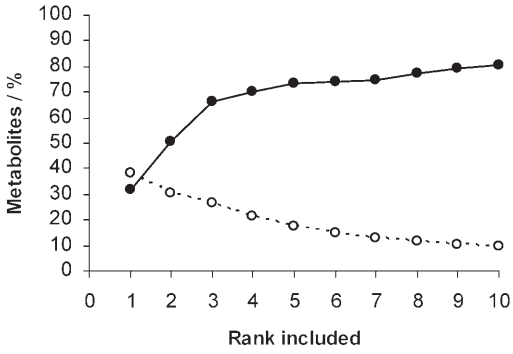

Figure 8. Percentage of P450 metabolites that are reproduced (recall; ——●——) and the percentage of the predictions that are experimentally confirmed (precision; -----○-----) as a function of the number of reactions from the top of the ranking list that is taken into account.

Comparison with other software

An extensive comparison of various metabolite prediction methods is beyond the scope of this paper. Nevertheless, the performance of SyGMa could be put into perspective by comparing it against published results from two well-known metabolite software packages: MetaSite and Meteor. The methodology used by MetaSite is based more on first principles and differs significantly from SyGMa. Furthermore, it predicts atomic sites in the molecule likely to be metabolized, rather then actual metabolites. MetaSite was evaluated by Zhou et al.[29] on the basis of a dataset of CYP 3A4-catalyzed metabolic reactions. This dataset was not reported and is likely to overlap significantly with our training set. Therefore, we did not test SyGMa with this dataset. Instead, it appears reasonable to assume that the challenge posed by the test set by Zhou et al. is comparable with our own test set of P450 reactions. MetaSite identified 60–70% of all metabolic sites in the top three of the ranking list[29] (figure 3 in ref. [29], top left panel), depending on the CYP 3A4 protein structure used. This is similar to the 66% of P450 metabolites that are reproduced by SyGMa within the top three (Figure 8). The present evaluation of SyGMa is in fact more stringent, because a SyGMa pre-

diction was considered correct only if the true metabolite was generated, whereas for MetaSite only the correct site of metabolism was sufficient.

Meteor, like SyGMa, is a rule-based method. Whereas SyGMa ranks metabolites on a practically continuous probability scale, Meteor assigns one out of five categories of likelihood to each predicted metabolite, which are designated Probable, Plausible, Equivocal, Doubted, and Improbable.[15] Testa et al.[16] evaluated Meteor results for 10 drugs, four of which were presented with enough detail to allow a comparison with SyGMa. The test involved single-step metabolic reactions only, and to evaluate the number of true predictions the first three categories, Probable, Plausible, and Equivocal, were considered positive. In all four cases the true predictions by Meteor included predictions assigned an "Equivocal" likelihood. All of the correctly predicted metabolites by Meteor were also generated by SyGMa. SyGMa does not define a threshold for positive and negative predictions. Therefore, to compare both approaches we consider the top *N* predictions by SyGMa as positive, with *N* chosen to include these common true predictions. Table 1 presents the results of the comparison. In summary: for galanthamine and indinavir, Meteor and SyGMa perform equally in terms of the number of true and false predictions. For the other two compounds SyGMa outperformed Meteor: in the case of tramadol, SyGMa identifies an additional observed metabolite that was missed by Meteor, and for omapatrilat SyGMa predicts the same number of observed metabolites among a much smaller number of false predictions.

Table 1. Comparison between SyGMa and Meteor predictions for four drugs.

Predictions	Meteor	SyGMa
galanthamine		
true predictions	6	6
false predictions	8	8
missed	0	0
tramadol		
true predictions	4	5
false predictions	5	2
missed	4	3
omapatrilat		
true predictions	2	2
false predictions	11	4
missed	1	1
indinavir		
true predictions	5	5
false predictions	31	30
missed	0	0

Species differences

The probabilities based on human in vivo data were compared with probabilities calculated on the basis of human in vitro data, rat in vivo data, and rat in vitro data (without changing the rules). Figure 9a indicates that overall, the probabilities obtained with human in vitro data correlate well with the probabilities based on the in vivo data. However, significant differences are present for some rules. Some of these differences can be rationalized on the basis of the experimental differences. For example, two "outliers" in Figure 9a are identified to be N-hydroxylation (A) and N-acetylation (B) of aromatic amine groups, as in anilines. These reactions are depicted in Figure 10. N-Acetylation (B) has a relatively high probability in vivo, while its probability in vitro is low. This can be explained by the fact that in vitro experiments (i.e. microsomal incubations) in general lack *N*-acetyl transferase activity. On the other hand, N-hydroxylation (A) has an intermediate probability in vitro, whereas its probability in vivo is low. Possibly, in the absence of the *N*-acetyl transferase activity in in vitro experiments, N-hydroxylation becomes a more important metabolic route for aromatic amines.

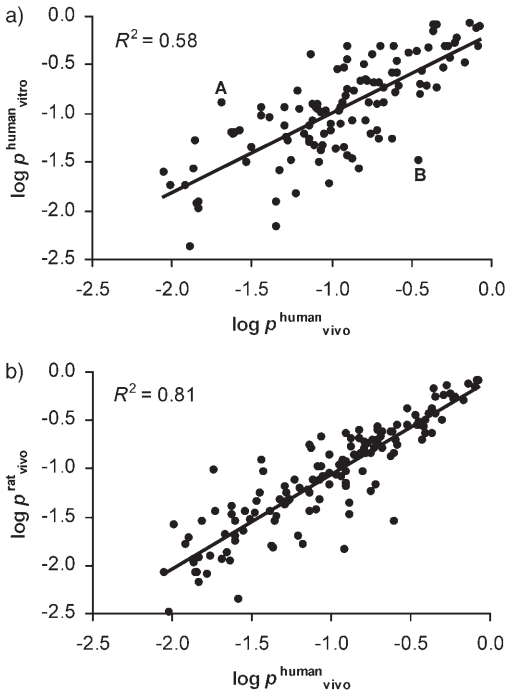

Figure 9. Probability scores for metabolic rules calculated based on a) human in vitro data and b) rat in vivo data plotted against probabilities based on human in vivo data. The data are presented on a logarithmic scale. Note that for 30 rules no examples were present in the human in vitro dataset and were left out in panel a). These are mostly phase 2 reaction rules, for example, for acetylation, sulfation, glycination, glucuronidation, and phosphorylation, but also some less common phase 1 oxidation reactions. Only three rules were not represented in the rat in vivo dataset and were left out in panel b).

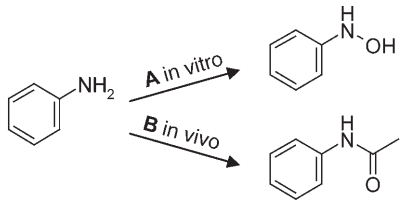

Figure 10. Metabolic routes of aromatic primary amines in vivo and in vitro.

Figure 9b shows that the correlation between probability scores from human and rat in vivo data is significantly higher. This indicates that interspecies differences between human and rat metabolism, in terms of overall probabilities for different types of reactions, are smaller than differences between in vivo and in vitro results.

Applications

To illustrate the use of SyGMa in the context of medicinal chemistry, it was applied to buspirone. The metabolism of buspirone has been extensively studied and involves many different metabolites. Figure 11 presents a summary of the metabolic pathways that are known from in vitro experiments, that is, in human liver microsomes,[30] indicated by the solid arrows. SyGMa reproduced all these experimental metabolites. Note that at ranks 3, 4, and 5, SyGMa generated the N-dealkylation products complementary to the 1-pyrimidinylpiperazine metabolite (rank 2) as well the two products from the alternative N-dealkylation reaction (cleaving off the 8-azaspiro(4,5)decane-7,9-dione moiety) at probabilities equal to rank 2. The metabolite reproduced at rank 221 involved three subsequent reactions steps, hydroxylation (rank 9) as well as hydrolysis of the 2,6-piperidinedione ring followed by ring closure.

At Organon, SyGMa predictions are now used to guide fast metabolite screening in in vitro samples from microsomal stability assays, by using multiple reaction monitoring (MRM). In this approach a triple quadrupole mass spectrometer is set up to specifically detect a predefined set of masses (e.g. from SyGMa predictions) and fragments. This MRM approach has been evaluated with buspirone as a test case. In addition to the known metabolites, SyGMa predicted an N,N-de-ethylation, breaking up the piperazine ring, as well as oxidation of the piperazine ring by the addition of a keto group. These metabolic reactions are indicated with dashed arrows in Figure 11. The resulting −26 and +14 mass metabolites were confirmed by in-house MRM mass detection and subsequently in full metabolite ID experiments. In parallel, data were published by Fandiño et al.,[31] also indicating the formation of these metabolites. These findings illustrate how the predictions by SyGMa support experimental metabolite ID and can lead to the identification of unforeseen metabolites that may otherwise be missed in fast screening approaches.

In the SyGMa predictions for buspirone, para-hydroxylation of the pyrimidine moiety was assigned highest probability. Blocking metabolically labile groups is a sensible approach to improve metabolic stability.[32] Thus, on the basis of such a prediction, medicinal chemists searching for a more stable ligand may decide to introduce a fluorine substituent on the para position to block the hydroxylation step. Indeed, this *p*-fluoropyrimidyl analogue of buspirone has been reported to show an increased half-life of greater than one order of magnitude in the presence of CYP 3A4, the most important enzyme in the metabolism of buspirone.[33]

As another example of the potential of SyGMa in lead optimization, Figure 12 shows the top-ranked prediction for delavirdine, an HIV-1 reverse transcriptase inhibitor. This N-dealkylation reaction was predicted with a high probability score of 0.38 and is indeed the major route of metabolism.[34] The half-life of the compound in microsomal incubation is 11 min.[35] In an effort to increase the metabolic stability, an analogue of this compound has been synthesized in which the *N*-isopropyl group is replaced by an ethoxy group; it showed an increased half-life of 47 min.[35] In agreement with this observation,

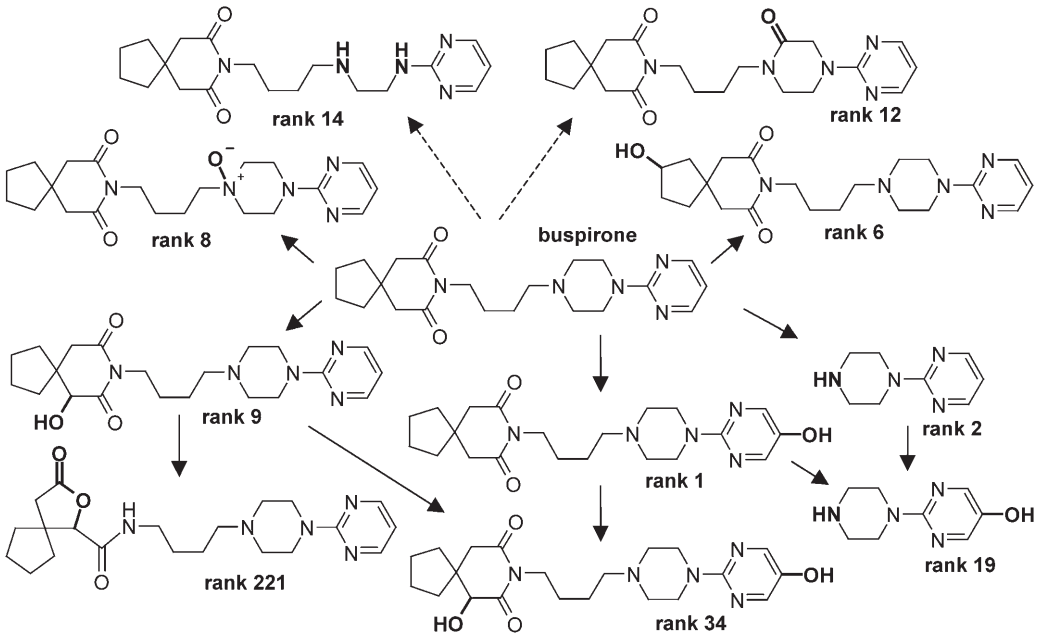

Figure 11. Metabolic pathways of buspirone as summarized by Zhu et al.,[30] complemented with two metabolites at the top, indicated by dashed arrows, which were predicted and confirmed experimentally in the present study. Rankings in the SyGMa list of predicted metabolites, using phase 1 rules only, are indicated. Reaction centers are indicated in boldface.

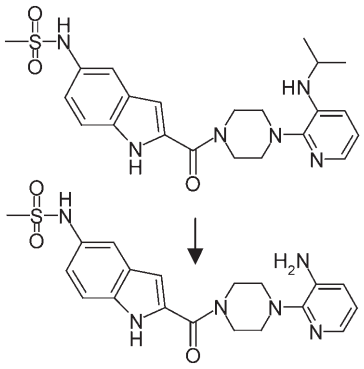

Figure 12. Predicted main metabolic reaction for delavirdine.

SyGMa predicts the analogous metabolic reaction, O-de-ethylation, to be much less likely, that is, with a probability score of 0.08. In fact, other metabolic reactions are predicted at higher probabilities for this analogue: aromatic hydroxylation, $p=$ 0.16, and hydrolysis of the amide bond, $p=0.10$. These corresponding metabolites have also been observed for delavirdine, however at lower quantities. The above examples illustrate that the predictions by SyGMa can be a useful tool to localize labile sites and to successfully guide medicinal chemistry aiming for metabolically more stable compounds.

Discussion

A rule-based method for the prediction of metabolites is presented that ranks predicted metabolites on the basis of probability scoring. Metabolic rules were developed by a combination of "expert knowledge" and empirical evaluation using a large database of experimentally observed metabolic reactions. A similarity analysis based on reaction fingerprints is presented that provides a 2-dimensional representation of the diversity in the reaction dataset, similar to another recently published approach based on self-organizing maps.[36] The latter approach used calculated properties of the bonds in reactants and products, whereas the present approach uses the chemical structure of the molecules directly. An example is provided of how this analysis can help to find gaps in a set of rules. Furthermore, a procedure is described to refine and optimize the rules in order to improve the overall performance of the rule base. About 70% of all known metabolic reactions observed in human studies, both in the training set and the test set, are covered by the current set of rules. Probability scores were calculated from the fraction of the predicted metabolites that are observed experimentally in the training set for each individual rule. Ranking of the predicted metabolites on the basis of the resulting probability scores is shown to successfully enrich the predictions with true metabolites. Overall, 30% of all observed metabolites in the test set, including 36% of the "major" metabolites, are in the top three ranked predictions; 45% of all observed metabolites in the test set, including 49% of the "major" metabolites, are in the top 10. The performance with a P450-specific test set is even better than with the complete test set. In total, 84% of the P450 metabolites were reproduced by SyGMa, 66% of the metabolites were predicted in the top three of the ranking list. The improved performance on the latter test set may be partly explained by the fact that it involves only single-step reactions. The prediction of multi-step metabolites, like with the complete dataset, is more challenging. Combinations of appropriate with inappropriate metabolic reactions result in false metabolites, even if the prediction was "partially correct". This effect does not occur in the P450 dataset. It is important to note, however, that the performance on the set of single-step reactions may be more representative for the application in a medicinal chemistry setting, in which the aim is to design metabolically stable compounds by blocking the first step in a metabolic pathway. Other reasons for the improved performance on the P450 dataset could be that it involves a more limited range of reaction mechanisms and that substrate–enzyme interactions (not taken into consideration by SyGMa) may have a smaller effect on the observed metabolism, as P450 enzymes are relatively substrate nonspecific. The results indicate the value of SyGMa for identifying labile sites in the search for chemical modifications that increase metabolic stability. Examples have been given that illustrate such application of SyGMa.

It is clear that not all possible metabolic reactions are currently covered in the rules. Rule-based systems inherently need continued effort in updating, as new data becomes available; novel chemical series may reveal new metabolic routes, and experimental methods to elucidate metabolites become increasingly powerful and may identify new metabolites not observed in the past. Analysis based on reaction fingerprints, as presented in this work, can help to maintain and update the rule base. However, it is also likely that a fraction of metabolic reactions may remain difficult to cover in general rules with a broader validity than single unique examples. The present approach requires sufficient examples for new rules in order to estimate a probability score, and this hampers the addition of rules on the basis of unique examples.

In the present approach, the effect of chemical environment on the susceptibility of a reaction center toward a certain reaction is accounted for by defining multiple rules covering different subsets of reaction centers. This approach takes chemical reactivity into account in an approximate and empirical way and, as mentioned above, is limited by the fact that each sub-rule requires sufficient examples. As an alternative, probabilities could be potentially correlated to calculated properties of the atomic reaction center. In a recently published approach, calculated atomic properties of reaction centers describing chemical reactivity were used to classify reactants from non-reactants.[37] Statistical techniques may be applied in a similar way to calculate probabilities suitable for ranking metabolites which could further enhance the present approach. It is important to note, however, that rule-based methods ignore the role of specific interactions and orientation in enzymes that catalyze the reaction. The question is whether a more sophisticated treatment of chemical reactivity would further improve the predictions as long as the influence of the metabolizing enzymes is not taken into account.

Given the demonstrated performance, SyGMa is a very suitable tool to quickly produce reasonable and fairly complete sets of potential metabolites. Such collections of potential metabolites are becoming increasingly valuable in experimental metabolite identification. Fast MRM or list-dependent MS^n screens can be setup on the basis of predicted metabolites to confirm their presence in in vitro or in vivo samples.[2,38] Furthermore, state-of-the-art software can import masses and structures of predicted metabolites and automatically confirm their presence in complex MS^E or MS–MS datasets collected in experimental metabolite identification studies.[3,38,39] In both ways, SyGMa has already proven its utility within experimental metabolite ID studies performed within Organon, resulting in the screening and identification of metabolites that would otherwise have been missed.[38,39]

Experimental Section

Datasets. Experimental metabolic reactions were retrieved from MDL's Metabolite Database, version 2001.[40] An advanced query was performed to retrieve only data from studies in humans and to exclude reactions with "presumed" reactants or products. Reactions labeled "optical resolution" were also excluded, as these reactions often refer to experimental analysis rather than actual metabolic processes and because the current approach does not take stereoisomerism into account. Furthermore, reactions with structures containing inorganic or undefined elements, such as –R or –X, were removed, as well as reactions involving large (non-druglike) molecules, that is, with molecular weight >900. The remaining dataset contained 6187 reactions observed with 1848 parent molecules. Reactions were labeled "Major" when they were annotated "Major" in the database on the basis of at least one referenced publication. From the 6187 reactions, the complete set of 3144 unique reactant structures was obtained which was used for the evaluation of reaction rules described below. The same procedure was followed for datasets of reactions observed in rat and reactions observed in in vitro studies using human and rat microsomes. Final evaluation of the method was performed with an independent test set, which was extracted from the update of the MDL Metabolite Database to the 2006 version, while the work on the metabolic rules was in progress. BCI fingerprints[41] were used to analyze the similarity between the test and training datasets. Each parent molecule in the test set was compared with the most similar compound in the training set. For 75% of the parent molecules this comparison yielded a Tanimoto coefficient <0.8, and for 50% <0.6, indicating sufficient diversity between the training and test sets. For the purpose of further evaluation, a subset of cytochrome P450 (CYP) reactions was taken from this new data, that is, reactions indicated to be metabolized by one or more CYP isoenzymes. Table 2 provides an overview of the various datasets used.

Table 2. Overview of the various datasets retrieved from the MDL Metabolite Database.

Dataset	Parents	Unique reactants	Reactions
Human in vivo	1848	3144	6187
Human in vitro	962	1270	2189
Rat in vivo	2765	4966	9262
Rat in vitro	1609	2205	3806
Human in vivo test-set	175	288	385
CYP test set	105	106	127

Reaction rules. We implemented a fast interpreter of generic reaction rules encoded in the Daylight SMIRKS language,[42] which allows systematic application of a set of rules to a compound structure for a specified number of subsequent steps in order to build up a complete reaction tree (as described in more detail below). A SMIRKS rule consists of a molecular substructure query (the "reactant side") and a definition of how the matching substructure is to be modified in the resulting product (the "product side"). Figure 1 provides some simple examples of reaction rules for primary alcohol oxidation. Atoms that are preserved in the reaction are matched between the reactant and product side by means of numeric labels (indicated by a colon). Unlabeled atoms on the reactant and product side disappear and appear, respectively. Furthermore, the SMIRKS language enables flexible query definitions, defining, for example, element, valency, aromaticity, charge, and ring membership of atoms as well as bond order and ring membership of bonds. This allows the definition of rules that apply to reaction centers with more or less specific chemical environments. Each rule was tested by applying it to all reactants in the dataset. The resulting products were compared with the metabolites reported in the database for the individual reactants. The number of generated metabolites that match the experimentally observed metabolites in the database was divided by the total number of metabolites generated (which is equal to the number of matches of the rule query in the reactant dataset). This ratio provides an empirical probability score for the metabolic rule:

$$p_{\text{rule}} = \frac{\text{number of correctly predicted metabolites}}{\text{total number of predicted metabolites}} \quad (1)$$

The reacting atom centers of the metabolic reactions matching a rule were examined with respect to the diversity of their direct chemical environment. When possible (see below), the rule was further refined to increase the probability ratio or split into multiple rules to account for differences in "reactivity" of different reaction centers toward the reaction. Refinement involved making the query part more specific, which results in a decrease in the number of incorrectly predicted metabolites. Division of rules resulted in multiple rules covering different subsets of the experimental reactions and yielding different probability ratios. For example, in the case of aromatic hydroxylations, the presence of ortho, meta, or para substituents were queried to distinguish more or less activated sites. For aliphatic reaction centers, relevant distinctions were made between reaction centers attached to aromatic, aliphatic, or heteroatomic cores. At the same time, rules were kept as general as possible to avoid over-fitting the rules to specific examples in the training set: each rule was required to match at least 10 compounds in the training set to yield a meaningful probability score, and the matching experimental reactions should cover multiple compound classes. Furthermore, rules with probability ratios below 0.01 were considered not predictive and either refined or rejected.

Predictions. Based on a complete set of rules, predictions are made by systematically applying all metabolic reactions to a parent molecule, thereby generating all possible metabolites. When the prediction is made for multiple subsequent reaction steps, each metabolite is again subjected to the set of reaction rules, and this is repeated for a preset maximum number of reaction steps. As a result, a network of metabolites connected by reactions is generated. When a single cleavage reaction results in multiple products, each product is treated as a separate metabolite. Me-

tabolites generated via more than one route are represented by a single "node" linking both branches of the metabolic network. This avoids duplication of metabolites as well as repetition of equivalent branches in the "metabolic tree". Minor cleavage products consisting of only a small fraction of the parent (e.g. resulting from hydrolysis or dealkylation of small groups) are often considered irrelevant. In the present implementation, small fragments are removed from the metabolic tree if they contain $<15\%$ of the atoms of the parent. This 15% cutoff was chosen based on the training set in which none of the experimental metabolites fell below this cutoff value.

When the metabolic network is completed, each metabolite is assigned the probability score from the reaction rule(s) from which it was formed. For metabolites resulting from multiple reaction steps, the probability scores of the individual steps are multiplied which assumes independence of the subsequent reactions of a metabolic pathway. This assumption is reasonable, as subsequent steps in a pathway are often carried out by different enzymes. When metabolites can be formed via different metabolic routes, the route resulting in the highest probability is selected. This approach was chosen to allow a rule set with overlapping (i.e. redundant) rules, for example, containing a general rule as well as a more specific rule covering a specific subset of reaction centers with a higher probability ratio. An example for which this approach is helpful is described in the Results section and concerns a redundancy in the rules for the oxidation of primary carbons. Finally, the 2D structures of all the metabolites in the reaction tree are generated. To facilitate visual inspection of the results, unchanged atoms of metabolites inherit the original coordinates, while coordinates for new atoms are optimized with respect to existing coordinates using a stochastic proximity embedding algorithm.[43] The metabolite structures are reported in order of decreasing probability score.

The predictions were evaluated in terms of recall and precision, defined as:

$$recall = \frac{\text{number of correctly predicted metabolites}}{\text{total number of experimental metabolites}} \quad (2)$$

$$precision = \frac{\text{number of correctly predicted metabolites}}{\text{total number of predicted metabolites}} \quad (3)$$

Note that Equation (3) for the precision is identical to Equation (1) for the probability of a rule. However, the precision is defined for the application of the entire rule set, whereas the probability score involves a singe rule.

Calculation of reaction fingerprints. An important feature of the rule base is its completeness in terms of coverage of the reactions in the training dataset. We implemented reaction fingerprints to analyze the contents of a reaction dataset. The fingerprints are used for clustering and visualization of the current training set, to analyze the coverage of the current rule base, and to support the search for new rules. The reaction fingerprints we applied describe the difference between the reactant and the product fingerprints and are based on an augmented atom description of the structures involved in a reaction.[44] First, fingerprints were generated for reactant and product molecules separately, based on Sybyl atom types and atom types augmented with a single layer around the central atom. This is illustrated in Table 3 on the basis of the example reaction in Figure 13. Up to 10 occurrences of an (augmented) atom type are distinguished. The difference fingerprint is defined by the differences in occurrence of each atom type in the reactant and product fingerprints. Thus, atom types with equal counts in the reactant and product fingerprints vanish in the difference fingerprint[45] (Table 4).

Table 3. Sybyl atom types and (1 layer) augmented atom types for propanol.[a]

Atom	1	2	3	4
Sybyl type	C.3	C.3	C.3	O.3
Augmented atom type	C.3 C.3	C.3 C.3 C.3	C.3 C.3 O.3	O.3 C.3

[a] Atom labels according to Figure 13.

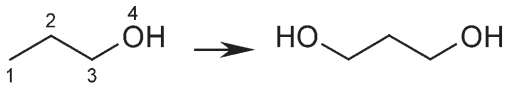

Figure 13. Reaction of propanol to propane-1,3-diol.

Table 4. Atomic fingerprints for the reactant 1-propanol and the product propane-1,3-diol, as well as the difference fingerprint to represent the reaction.

	C.3	O.3	C.3 C.3	C.3 C.3 C.3	C.3 C.3 O.3	O.3 C.3
Reactant	3	1	1	1	1	1
Product	3	2	0	1	2	2
Reaction	0	+1	−1	0	+1	+1

Based on the difference fingerprints, similarity coefficients such as the Tanimoto coefficient can be calculated between pairs of reactions and subsequently used for clustering or other types of analysis. Reactions which involve removal, addition, or modification of defined molecular groups have very similar fingerprints. It should be noted, however, that other reactions, such as dealkylation or hydrolysis, involve removal of nonspecific parts of a molecule, which may result in more different fingerprints. We used the calculated Soergel distance (i.e. 1.0 minus the Tanimoto similarity coefficient) in combination with a 2D projection based on stochastic proximity embedding[43] to visualize the contents of the reaction database and the coverage by the SyGMa rules. This method optimizes the distances between points on a 2D plane to correspond as much as possible to the distances calculated in the fingerprint space between all pairs of metabolic reactions. The resulting scatter plot provides a 2D map of the metabolic reactions in which similar reactions are clustered together.

Acknowledgements

We thank Jos Lommerse for helpful discussions and Peter Jacobs and Harrie Peters for the fruitful collaboration to integrate metabolite prediction and experimental metabolite identification within Organon.

Keywords: computer chemistry · drug design · metabolism · metabolite prediction · reaction fingerprints

[1] A. E. Nassar, R. E. Talaat, *Drug Discovery Today* **2004**, *9*, 317–327.
[2] M. R. Anari, R. I. Sanchez, R. Bakhtiar, R. B. Franklin, T. A. Baillie, *Anal. Chem.* **2004**, *76*, 823–832.
[3] C. C. Hao, S. Campbell, D. Stranz, N. McSweeney, "Identification of In-Vitro Metabolites of Indinavir using Automated LC/MS/MS Acquisition, In-Silico Prediction, and Structure-Based Data Analysis", *Proceedings of the 52nd ASMS Conference* **2004**, Nashville (USA).
[4] J. P. Jones, M. Mysinger, K. R. Korzekwa, *Drug Metab. Dispos.* **2002**, *30*, 7–12.
[5] S. B. Singh, L. Q. Shen, M. J. Walker, R. P. Sheridan, *J. Med. Chem.* **2003**, *46*, 1330–1336.
[6] J. L. Lewin, C. J. Cramer, *Mol. Pharmaceutics* **2004**, *1*, 128–135.
[7] I. M. C. M. Rietjens, A. E. M. F. Soffers, C. Veeger, J. Vervoort, *Biochemistry* **1993**, *32*, 4801–4812.
[8] M. E. Beck, *J. Chem. Inf. Model.* **2005**, *45*, 273–282.
[9] C. de Graaf, C. Oostenbrink, P. H. Keizers, T. van der Wijst, A. Jongejan, N. P. Vermeulen, *J. Med. Chem.* **2006**, *49*, 2417–2430.
[10] I. Zamora, L. Afzelius, G. Cruciani, *J. Med. Chem.* **2003**, *46*, 2313–2324.
[11] G. Cruciani, E. Carosati, B. B. De, K. Ethirajulu, C. Mackie, T. Howe, R. Vianello, *J. Med. Chem.* **2005**, *48*, 6970–6979.
[12] G. Klopman, M. Dimayuga, J. Talafous, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1320–1325.
[13] J. Talafous, L. M. Sayre, J. J. Mieyal, G. Klopman, *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1326–1333.
[14] F. Darvas, *J. Mol. Graphics* **1988**, *6*, 80–86.
[15] W. G. Button, P. N. Judson, A. Long, J. D. Vessey, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1371–1377.
[16] B. Testa, A. L. Balmat, A. Long, P. Judson, *Chem. Biodiversity* **2005**, *2*, 872–885.
[17] D. Korolev, K. V. Balakin, Y. Nikolsky, E. Kirillov, Y. A. Ivanenkov, N. P. Savchuk, A. A. Ivashchenko, T. Nikolskaya, *J. Med. Chem.* **2003**, *46*, 3631–3643.
[18] S. Ekins, S. Andreyev, A. Ryabov, E. Kirillov, E. A. Rakhmatulin, A. Bugrim, T. Nikolskaya, *Expert Opin. Drug Metab. Toxicol.* **2005**, *1*, 303–324.
[19] S. Ekins, S. Andreyev, A. Ryabov, E. Kirillov, E. A. Rakhmatulin, S. Sorokina, A. Bugrim, T. Nikolskaya, *Drug Metab. Dispos.* **2006**, *34*, 495–503.
[20] G. M. Banik, *Curr. Drug Discovery* **2004**, 31–34.
[21] G. Klopman, M. Tu, J. Talafous, *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 329–334.
[22] Y. Borodina, A. Sadym, D. Filimonov, V. Blinova, A. Dmitriev, V. Poroikov, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1636–1646.
[23] S. Boyer, C. H. Arnby, L. Carlsson, J. Smith, V. Stein, R. C. Glen, *J. Chem. Inf. Model.* **2007**, *47*, 583–590.
[24] O. G. Mekenyan, S. D. Dimitrov, T. S. Pavlov, G. D. Veith, *Curr. Pharm. Des.* **2004**, *10*, 1273–1293.
[25] S. Boyer, I. Zamora, *J. Comput.-Aided Mol. Des.* **2002**, *16*, 403–413.
[26] H. Yin, G. Bennett, J. P. Jones, *Chem.-Biol. Interact.* **1994**, *90*, 47–58.
[27] The SyGMa rule base is available on request from the authors.
[28] J. B. Mangold, H. Gu, L. C. Rodriguez, J. Bonner, J. Dickson, C. Rordorf, *Drug Metab. Dispos.* **2004**, *32*, 566–571.
[29] D. Zhou, L. Afzelius, S. W. Grimm, T. B. Andersson, R. J. Zauhar, I. Zamora, *Drug Metab. Dispos.* **2006**, *34*, 976–983.
[30] M. Zhu, W. Zhao, H. Jimenez, D. Zhang, S. Yeola, R. Dai, N. Vachharajani, J. Mitroka, *Drug Metab. Dispos.* **2005**, *33*, 500–507.
[31] A. S. Fandiño, E. Nagele, P. D. Perkins, *J. Mass Spectrom.* **2006**, *41*, 248–255.
[32] A. E. Nassar, A. M. Kamel, C. Clarimont, *Drug Discovery Today* **2004**, *9*, 1020–1028.
[33] M. Tandon, M. M. O'Donnell, A. Porte, D. Vensel, D. Yang, R. Palma, A. Beresford, M. A. Ashwell, *Bioorg. Med. Chem. Lett.* **2004**, *14*, 1709–1712.
[34] R. L. Voorman, S. M. Maio, M. J. Hauer, P. E. Sanders, N. A. Payne, M. J. Ackland, *Drug Metab. Dispos.* **1998**, *26*, 631–639.
[35] M. J. Genin, T. J. Poel, Y. Yagi, C. Biles, I. Althaus, B. J. Keiser, L. A. Kopta, J. M. Friis, F. Reusser, W. J. Adams, R. A. Olmsted, R. L. Voorman, R. C. Thomas, D. L. Romero, *J. Med. Chem.* **1996**, *39*, 5267–5275.
[36] D. A. Latino, J. Aires-de-Sousa, *Angew. Chem.* **2006**, *118*, 2120–2123; *Angew. Chem. Int. Ed.* **2006**, *45*, 2066–2069.
[37] F. Mu, P. J. Unkefer, C. J. Unkefer, W. S. Hlavacek, *Bioinformatics* **2006**, *22*, 3082–3088.
[38] P. L. Jacobs, H. A. M. Peters, L. Ridder, M. Wagener, "Metabolite Prediction And A Two-Step Metabolite Identification Approach in Early Drug Discovery", *Proceedings of the 54th ASMS Conference* **2006**, Seattle (USA).
[39] P. L. Jacobs, E. Meulen, L. Ridder, M. Wagener, "Metabolite Prediction and Accurate Mass UPLC–MS^E in Metabolite Identification", *Proceedings of the 55th ASMS Conference* **2007**, Indianapolis (USA).
[40] MDL Metabolite Database, Elsevier (2001): http://www.mdl.com/products/predictive/metabolite.
[41] J. M. Barnard, G. M. Downs, *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 644–649.
[42] Daylight Chemical Information Systems, Inc. **2006**, http://www.daylight.com/dayhtml/doc/theory/index.html (accessed January 31, 2008).
[43] D. K. Agrafiotis, H. Xu, *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 15869–15872.
[44] A. Bender, H. Y. Mussa, R. C. Glen, S. Reiling, *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708–1718.
[45] Q. Y. Zhang, J. Aires-de-Sousa, *J. Chem. Inf. Model.* **2005**, *45*, 1775–1783.

Received: November 2, 2007
Revised: January 23, 2008
Published online on February 29, 2008

