BIOINFORMATICS ORIGINAL PAPER
Vol. 24 no. 18 2008, pages 2079–2085
doi:10.1093/bioinformatics/btn378

Data and text mining

Data-driven extraction of relative reasoning rules to limit combinatorial explosion in biodegradation pathway prediction

Kathrin Fenner[1,2,3,*], Junfeng Gao[4], Stefan Kramer[5], Lynda Ellis[4] and Larry Wackett[3]

[1]Eawag, Swiss Federal Institute of Aquatic Science and Technology, CH-8600 Dübendorf, [2]Institute of Biogeochemistry and Pollutant Dynamics (IBP), ETH Zurich, CH-8092 Zürich, Switzerland, [3]Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, St. Paul, MN 55108, [4]Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN 55455, USA and [5]Institute for Informatics/I12, Technical University of Munich, Boltzmannstr. 3, D-85748 Garching bei München, Germany

Received on October 16, 2007; revised on June 19, 2008; accepted on July 17, 2008

Advance Access publication July 19, 2008

Associate Editor: Thomas Lengauer

ABSTRACT

Motivation: The University of Minnesota Pathway Prediction System (UM-PPS) is a rule-based expert system to predict plausible biodegradation pathways for organic compounds. However, iterative application of these rules to generate biodegradation pathways leads to combinatorial explosion. We use data from known biotransformation pathways to rationally determine biotransformation priorities (relative reasoning rules) to limit this explosion.

Results: A total of 112 relative reasoning rules were identified and implemented. In one prediction step, i.e. as per one generation predicted, the use of relative reasoning decreases the predicted biotransformations by over 25% for 50 compounds used to generate the rules and by about 15% for an external validation set of 47 xenobiotics, including pesticides, biocides and pharmaceuticals. The percentage of correctly predicted, experimentally known products remains at 75% when relative reasoning is used. The set of relative reasoning rules identified, therefore, effectively reduces the number of predicted transformation products without compromising the quality of the predictions.

Availability: The UM-PPS server is freely available on the web to all users at the time of submission of this manuscript and will be available following publication at http://umbbd.msi.umn.edu/predict/.

Contact: kathrin.fenner@eawag.ch

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Persistence of man-made chemicals (xenobiotics) in the environment is largely predicated on their susceptibility to degradation by microbial metabolism. Thus, a thorough understanding of microbial degradative metabolism is a crucial component in environmental risk assessment of chemicals. In this context, more data on microbial degradation is now mandated by new legislation in Europe concerning industrial chemicals (REACH, 2006) and human and veterinary medicines (EMEA, 2006; VICH, 2004).

Due to these increased data requirements, *in silico* methods to predict biodegradation behavior at an early stage of environmental risk assessment are increasingly sought. So far, these tools are mostly geared towards predicting general biodegradability; for example, whether or not a chemical passes a regulatory ready biodegradability test (Jaworska *et al.*, 2003). Most recently, more focus has been directed toward predicting biodegradation pathways and products that may accumulate in the environment. Regulatory requirements to determine stable transformation products have been driven by multiple reports on transformation products being found at higher concentrations than the initial parent compounds (Battaglin *et al.*, 2005; Kolpin *et al.*, 2004). Anticipating what products might be formed with untested chemicals is important at several levels. It can identify potentially toxic and stable transformation products at the screening stage of environmental risk assessment. Moreover, products predicted *in silico* can be used to guide chemical analysis in degradation studies. Current tools that predict biodegradation pathways include META (Klopman and Tu, 1997), CATABOL (Jaworska *et al.*, 2002) and the UM-PPS (Hou *et al.*, 2004). They all belong to the category of artificial intelligence systems, in that they are based on a set of transformation rules that recognize compound substructures and transform them into product substructures according to these rules.

The most recently created, the University of Minnesota Pathway Prediction System (UM-PPS) (Hou *et al.*, 2004), uses substructure searching and atom-to-atom mapping to transform a query substrate into a product when a rule's substrate substructure is present in the query. The UM-PPS biotransformation rules (btrules) are primarily based on the University of Minnesota Biodegradation/Biocatalysis Database (UM-BBD) (Ellis *et al.*, 2006), a manually curated collection of almost 200 microbially mediated metabolic pathways, consisting of over 1000 enzyme-catalyzed reactions. The set of about 200 UM-PPS rules covers over 90% of appropriate reactions in the UM-BBD. However, simple application of all possible rules typically yields many predicted transformation products, some of which may not occur in nature and thus are false positives. False positives lead to combinatorial explosion when the rules are

*To whom correspondence should be addressed.

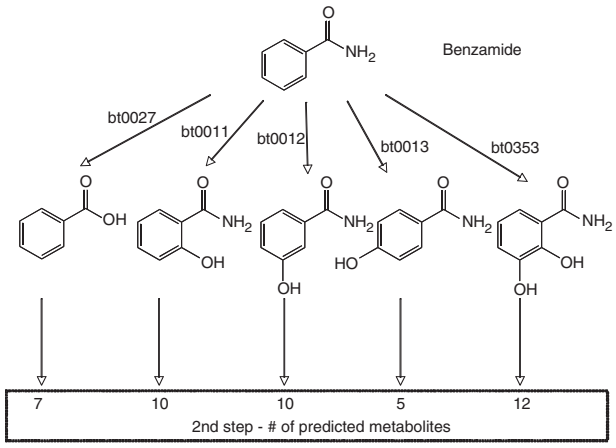

Fig. 1. Combinatorial explosion over the first two generations of predicted benzamide biotransformations with UM-PPS.

iteratively applied to predict consecutive steps of biodegradation reactions. Even a structurally simple compound may trigger multiple transformation rules and may thus yield many predicted products as illustrated in Figure 1. An even greater number of products may be observed for *in silico* predictions with environmentally relevant, xenobiotic chemicals, which frequently contain multiple functional groups. This problem is not exclusive to UM-PPS, but applies in principle to all rule-based artificial intelligence systems for pathway prediction. Note that this type of combinatorial explosion is different from the one encountered in the construction and analysis of large metabolic flux networks (e.g. Urbanczik and Wagner, 2005).

Others use various methods to limit biodegradation pathway prediction: (i) restriction of the rules' applicability domain through further specifications in terms of molecular descriptors (Embrechts and Ekins, 2007; Mu *et al*., 2006), molecular substructures (Gomez *et al.*, 2007) or chemical similarity (Oh *et al.*, 2007), or (ii) predefinition of rule probabilities modeled on biological oxygen demand (BOD) data (Jaworska *et al.*, 2002). These methods exhibit different shortcomings that limit their applicability for this task, including the need for defining negative prediction outcomes to train statistical methods based on molecular descriptors, restriction of the analysis to the chemical environment immediately adjacent to the functional group being transformed (Mu *et al*., 2006), the inappropriateness of chemical similarity measures to describe biochemical similarity and the dichotomy between coverage and prediction accuracy in the similarity approach (Oh *et al.*, 2007).

In an initial effort to deal with combinatorial explosion in UM-PPS, expert knowledge was provided to rank UM-PPS rules into five aerobic likelihood groups (very likely, likely, neutral, unlikely and very unlikely) (Ellis *et al*., 2006). This 'absolute' reasoning approach can be used to reduce the number of aerobic predictions, e.g. by removing unlikely and very unlikely biotransformations. However, since only 20% of the UM-PPS btrules have an aerobic likelihood of unlikely or very unlikely, this approach alone is insufficient.

In the present study, we explore the complementary approach of using relative reasoning. Relative reasoning has previously been used to limit combinatorial explosion in the prediction of mammalian detoxification pathways (Button *et al.*, 2003). Our approach accounts for the fact that there are almost invariably several functional groups in a molecule that might potentially react. In this case, it gives priority to those of a set of applicable rules that, based on existing knowledge of biotranformation pathways, are deemed to encode preferred biotransformations. Knowledge on which biotransformations for a specific molecular structure are preferred is extracted from the UM-BBD database. While the goal of this project is to restrict the number of predicted products, the new system should not be overly restrictive since most xenobiotics are degraded by multiple metabolic pathways in the environment.

2 METHODS

2.1 Extraction of rule priorities from UM-BBD

Information on rules priorities was extracted from the over 1000 reactions reported in the UM-BBD database and stored in a matrix. On July 24, 2007 the UM-PPS contained 204 btrules and the UM-BBD contained 1084 compounds. Of these, 366 compounds not associated with any rule (terminal compounds of reported pathways, compounds containing metals or other compounds whose biodegradation should not be predicted, http://umbbd.msi.umn.edu/predict/notbepredicted.html) were removed. Likewise 25 strictly anaerobic (unlikely or very unlikely) btrules and btrules not triggered by any compound in the UM-BBD were removed. The remaining 718 UM-BBD compounds were submitted to 179 UM-PPS btrules. Three outcomes were distinguished and encoded into a 718×179 matrix D, with the columns representing all UM-PPS btrules and the rows representing all UM-BBD compounds. The outcomes for the 64 261 ($718 \times 179/2$) possible compound/rule combinations were distinguished as follows.

- Compound c_i does not trigger a given btrule r_j (encoded as $D_{ij} = -1$).
- Compound c_i triggers a given btrule r_j, but the known reaction(s) for that compound reported in UM-BBD proceed(s) according to a different btrule (encoded as $D_{ij} = 0$).
- Compound c_i does trigger a given btrule r_j, and the transformation encoded by that btrule represents a known reaction for that compound in UM-BBD (encoded as $D_{ij} = 1$).

The resulting matrix, which can be obtained from the authors on request, contains all information on possible and known reactions of the 718 UM-BBD compounds.

This information was further analyzed to find relative priorities between pairs of rules, which could then be used as relative reasoning rules in UM-PPS.

2.2 Identification of relative reasoning rules

To identify valid relative reasoning rules, the matrix was searched for pairs of rules with compounds in common (i.e. compounds that trigger both rules). These rule pairs were further examined to find those for which the known pathways for the compounds held in common (i.e. compounds that triggered both rules) all proceed exclusively according to one of the two rules. Such rule pairs are candidates for relative reasoning rules and are referred to as one-directional rule pairs.

A one-directional relationship with five common compounds shared by two rules has only a 3% probability to occur ($0.5^5 = 0.03125$). Therefore, it could be regarded as sufficiently robust to be considered a valid relative reasoning rule. However, in chemical terms, diversity of molecular structures is another prerequisite to define broadly applicable relative reasoning rules. Structural diversity is given if the functional groups that trigger the two rules of a rule pair are present in several different stereo- and electrochemical arrangements in the set of common compounds. This prerequisite is not always fulfilled for the compounds within the UM-BBD. We therefore defined two levels of relative reasoning rules.

Table 1. Reaction groups and btrules assigned to each

Reaction group	btrules
Alcohol oxidation	bt0001, bt0002
Aldehyde oxidation	bt0003
Aliphatic hydroxylation	bt0241, bt0242, *bt0036, bt0332, bt0333, bt0334*
Amide hydrolysis	bt0067, bt0318, *bt0024, bt0027*
Aromatic *vic*-diol ring cleavage	bt0008, bt0254, *bt0041, bt0045, bt0069, bt0131, bt0165, bt0174, bt0184, bt0297*
Aromatic ring dioxygenation	bt0005, *bt0042, bt0055, bt0065, bt0072, bt0128, bt0196*
Aromatic ring monooxygenation	bt0011, bt0012, bt0013
Phenolic ring monooxygenation	bt0014, bt0064
C=C bond reactions	bt0021, bt0049, bt0259, bt0291
CoA-thioester formation	bt0094, *bt0315*
Decarboxylation	bt0051, bt0082, *bt0060, bt0072*
Keto-ene hydrolysis	bt0040, bt0047
Keto-enol tautomerism	bt0231, *bt0044*
Oxidation of *vic*-di-H-di-OH to aromatic	bt0255, *bt0056, bt0197*

Rules in normal font are part of a strict relative reasoning rule; rules in italics were added to develop the extended rule set (see text). Details of each btrule, for example, bt0001, can be seen using the URL: http://umbbd.msi.umn.edu/servlets/rule.jsp?rule=bt0001

For the first, strict level, a conservative criterion of a minimum of 10 common compounds showing a one-directional relationship was used to select relative reasoning rules. To be more precise, let T_i be the set of compounds for which rule r_i is triggered (i.e. $T_i = \{j | D_{ji} = 0 \vee D_{ji} = 1\}$ and O_i be the set of compounds for which rule r_i is triggered *and* represents a known UM-BBD reaction for that compound (i.e. $O_i = \{j | D_{ji} = 1\}$, thus, $O_i \subseteq T_i$). (Note that clearly $O_i \cap T_i \cap T_j = O_i \cap T_j$ and, vice versa, $T_i \cap T_j \cap O_j = T_i \cap O_j$.) The condition for the selection of strict relative reasoning rules then reads as:

$$R'_A = \{(r_i > r_j) \mid |O_i \cap T_j| \geq 10 \wedge |T_i \cap O_j| = 0\}$$

R'_A was post-processed by inspecting the compound structures within each set of common compounds to confirm diversity of the reaction centers and their relative positions in the molecules. Since scientific knowledge on enzyme specificities for a broad set of enzymes is clearly insufficient to derive objective measures of the required structural diversity, it was dealt with as follows. If structural diversity was limited, i.e. three or less molecular substructures were sufficient to represent the relative positions of the reaction centers in the set of common compounds, the relative reasoning rule was restricted to only apply to compounds containing those exact substructures (for an example, see Section 3.1). This yielded a final set of strict relative reasoning rules, R_A.

For the second, extended level, the search for relative reasoning rules was extended using analogy reasoning. For each btrule that was part of a relative reasoning rule on the first level, additional btrules that belong to the same type of reaction were identified (reaction groups, see Table 1). For each strict relative reasoning rule, all possible pairings of btrules belonging to the two respective reaction groups were then analyzed (reaction group matrix). If there was no contradiction to the one-directional relationship of the strict relative reasoning rule across the entire reaction group matrix, rule pairs within that matrix that exhibited at least five common compounds or with at least one common compound but for which one of the two btrules was part of the corresponding strict rule pair were included into the set of extended (second level) relative reasoning rules. In mathematical terms, let group(r_i) denote the set of rules that belong to the same type of reaction as r_i. The condition for the selection of additional relative reasoning rules to be included in the extended set then reads as:

$$R_B = \{(r_i > r_j) | (r'_i > r'_j) \in R_A \wedge r_i \in \text{group}(r'_i) \wedge r_j \in \text{group}(r'_j) \wedge |T_i \cap T_j| \geq 1 \wedge (\neg\exists c\colon c \notin O_i \wedge c \in O_j) \wedge ((|O_i \cap T_j| \geq 5 \wedge |T_i \cap O_j| = 0) \vee r_i = r'_i \vee r_j = r'_j)\}$$

All proposed relative reasoning rules were found to conform to common knowledge of microbial metabolism. It was also tested whether there were any contradictions in priorities across all relative reasoning rules. No such contradictions were found despite relative reasoning rules being derived by pairwise analysis. These two analyses confirmed that our procedure was sufficiently stringent to avoid chance relationships and produced results representing valid metabolic logic.

2.3 Validation method

2.3.1 Testing sets Two sets of compounds were selected for validation of the relative reasoning approach. The first set consisted of 50 randomly selected UM-BBD compounds. This set contained both starting compounds and intermediary metabolites of UM-BBD pathways. Since most of these compounds were part of the training set, the main purpose of this set was to check the correctness of our procedure to identify and implement relative reasoning rules. The second, external validation set contained 47 xenobiotic compounds from different chemical classes: 24 pesticides, 7 biocides and 16 pharmaceuticals (compound names, CAS numbers and molecular structures are given in Table S1 of the Supplementary Material).

2.3.2 Validation method Performance of the relative reasoning approach was measured in comparison to the performance of UM-PPS before the implementation of relative reasoning. The evaluation was done in two stages, first implementing the set of strict relative reasoning rules, and second implementing the set of extended relative reasoning rules. Three measures were defined to quantify the effect of implementing relative reasoning rules. To calculate them, we counted the number of predicted transformation reactions *PR*, the number of known reactions that are correctly predicted KR_p and the number of known reactions that are not predicted KR_{np}. These outcomes were evaluated for prediction of the first generation of transformation products, i.e. for one transformation step only. The last two outcomes, KR_p and KR_{np}, were evaluated for the UM-BBD validation set and for 25 pesticides and biocides from the xenobiotics validation set. For the UM-BBD validation set, the known reactions were available from the UM-BBD database. For the xenobiotics validation set we used the handbooks of Roberts (1998) on metabolic pathways for pesticides, and the scientific literature, to identify reactions leading to known first generation transformation products (the experimentally known biodegradation products, including names, CAS numbers and molecular structures are given in Table S2 of the Supplementary Material).

The three performance measures defined are reduction (Equation (1)), indicating by what percentage the number of predicted transformation reactions could be reduced by implementing relative reasoning; sensitivity (Equation (2)), indicating what percentage of known transformation reactions are captured by the UM-PPS predictions; and selectivity (Equation (3)), a measure of prediction stringency, indicating how many of the predicted reactions correspond to known products.

$$\text{Reduction}: \quad \left(1 - \frac{PR_{\text{ARR}}}{PR_{\text{BRR}}}\right) \cdot 100 \tag{1}$$

$$\text{Sensitivity}: \quad \left(\frac{KR_p}{KR_p + KR_{np}}\right) \cdot 100 \tag{2}$$

$$\text{Selectivity}: \quad \frac{KR_p}{PR} \cdot 100 \tag{3}$$

PR_{BRR} is the number of predicted reactions before introduction of relative reasoning, and PR_{ARR} is the number of predicted reactions after introduction of relative reasoning.

Table 2. Relative priorities among reaction groups

Reaction Group Name (abbr.)	Priority over
Alcohol oxidation (acx)	arm, alh
Aldehyde oxidation (adx)	ard, arm, cc, alh
Aliphatic hydroxylation (alh)	–
Amide hydrolysis (amh)	alh
Aromatic *vic*-diol ring cleavage (arc)	ard, arm
Aromatic ring dioxygenation (ard)	–
Aromatic ring monooxygenation (arm)	ard
Phenolic ring monooxygenation (prm)	arm
C=C bond reactions (cc)	alh
CoA-thioester formation (coa)	alh
Decarboxylation (dc)	–
Keto-ene hydrolysis (keh)	ard, arm, cc, coa, dc, ket
Keto-enol tautomerism (ket)	–
Oxidation of *vic*-di-H-di-OH to aromatic (vda)	acx, ard, arm, cc

3 RESULTS AND DISCUSSION

3.1 Relative reasoning rules

Of the 90 rule pairs with $\geq$10 compounds in common, 46 were clearly one-directional. Inspection of the structural diversity of the common compounds of each one-directional rule pair led to the restriction of the applicability domain of one rule pair (bt0067 > bt0242) to secondary amides only. The other 45 one-directional relationships were used directly as strict relative reasoning rules. The btrules involved in the strict relative reasoning rules could be attributed to 14 different reaction groups. Table 1 lists the different reaction groups and the btrules assigned to them. It distinguishes between btrules that are part of strict relative reasoning rules and btrules that were assigned to the same reaction groups for use in analogy reasoning. Table 2 indicates the priorities between reaction groups as derived from the strict rule set. Based on the reaction groups and priorities in Tables 1 and 2, an additional 66 relative reasoning rules were identified for the extended set of rules. Table 3 gives the two sets of strict and extended relative reasoning rules thus derived.

3.2 Implementation

The relative reasoning rules were implemented in three different ways. For two of the biotransformations, oxidation of aldehyde to carboxylate (bt0003) and oxidation of *vic*-dihydrodihydroxyaromatic to *vic*-dihydroxyaromatic (bt0255), evidence from gene sequencing and enzymatic studies suggest that they proceed very readily once their substrate substructures (aldehyde for bt0003 or *vic*-dihydrodihydroxyaromatic for bt0255) are formed. The high likelihoods of bt0003 and bt0255 were further confirmed by inspection of the extracted rule priority matrix, which showed that in 211 out of 215 (for bt0003) and in 237 out of 238 cases (for bt0255) these btrules have priority over the other applicable btrules. These rules were, therefore, implemented as so-called immediate btrules in UM-PPS. For immediate btrules only the product of the immediate btrule (carboxylate for bt0003 or *vic*-dihydroxyaromatic for bt0255) is shown to the user; the user is not given the choice of selecting any of the other theoretically possible transformation products of the starting compound.

Table 3. Strict (A) and extended (B) relative reasoning rules derived from the extracted rule priorities

A. Strict relative reasoning rules

bt0001 > bt0011, bt0012
bt0002 > bt0241, bt0242
bt0003[a] > bt0005, bt0011, bt0012, bt0013, bt0021, bt0049, bt0291, bt0242
bt0008 > bt0005, bt0011, bt0012, bt0013, bt0014, bt0064
bt0014 > bt0005
bt0021 > bt0242
bt0040 > bt0049, bt0231
bt0047[a] > bt0005, bt0011, bt0012, bt0021, bt0049, bt0291, bt0040, bt0051, bt0082, bt0094
bt0067 > bt0242[b]
bt0094 > bt0241, bt0242
bt0254 > bt0005, bt0014, bt0064
bt0255[a] > bt0002, bt0005, bt0011, bt0012, bt0013, bt0021, bt0049, bt0291

B. Extended relative reasoning rules

bt0001 > bt0011, bt0012, bt0013, bt0014, bt0064, bt0241, bt0242
bt0002 > bt0011, bt0012, bt0013, bt0241, bt0242, bt0332, bt0333, bt0334
bt0008 > bt0005, bt0011, bt0012, bt0013, bt0014, bt0055, bt0064, bt0065, bt0128
bt0014 > bt0005, bt0011[c], bt0012[c], bt0013[c]
bt0021 > bt0241, bt0242
bt0040 > bt0005, bt0011, bt0012, bt0049, bt0231
bt0041 > bt0005, bt0011, bt0012, bt0013, bt0014, bt0064
bt0045 > bt0005, bt0011, bt0012, bt0013, bt0014, bt0064
bt0056 > bt0002, bt0005, bt0011, bt0012, bt0013, bt0021, bt0049, bt0291
bt0067 > bt0242[b]
bt0069 > bt0014
bt0094 > bt0241, bt0242
bt0131 > bt0005, bt0014, bt0064
bt0165 > bt0005, bt0014, bt0064
bt0174 > bt0005, bt0011, bt0012, bt0013, bt0014
bt0197 > bt0002, bt0005, bt0011, bt0012, bt0013, bt0021, bt0049, bt0291
bt0254 > bt0005, bt0011, bt0012, bt0014, bt0055, bt0064, bt0128
bt0297 > bt0005

[a]These strict rules were treated separately and do not form part of the extended rule set (see text).
[b]This relative reasoning rule was only applied to secondary amides (see text).
[c]There were only 8, 7 and 6 common compounds between bt0014 and bt0011, bt0012 and bt0013, respectively, and no corresponding strict relative reasoning rule. However, since these rule priorities represent common metabolic logic, they were nonetheless implemented at the extended level.

Another btrule, bt0047 (hydrolytic cleavage of 2-oxo-3-enoate-4-aryl compounds to pyruvate and aromatic aldehydes), was also identified as a candidate immediate btrule: in 2 out of 175 cases involving bt0047 as one btrule of a rule pair, both bt0047 and the other btrule represent known reactions for their common compound. In all remaining 173 cases bt0047 has priority. However, because bt0047 is an integral part of a common pathway for the degradation of condensed aromatic compounds, it was not implemented as an immediate btrule. Instead, the entire pathway was implemented as a single super rule covering a series of reactions in a fixed sequence, with bt0047 included as one step of that super rule. The overall effect of this on the number of predicted reactions is the same as that of an immediate btrule.

All other relative reasoning rules identified in Table 3 were implemented such that whenever two btrules for which a relative reasoning rule exists are triggered for a given compound, only the product of the btrule with the higher priority is shown. Other rules

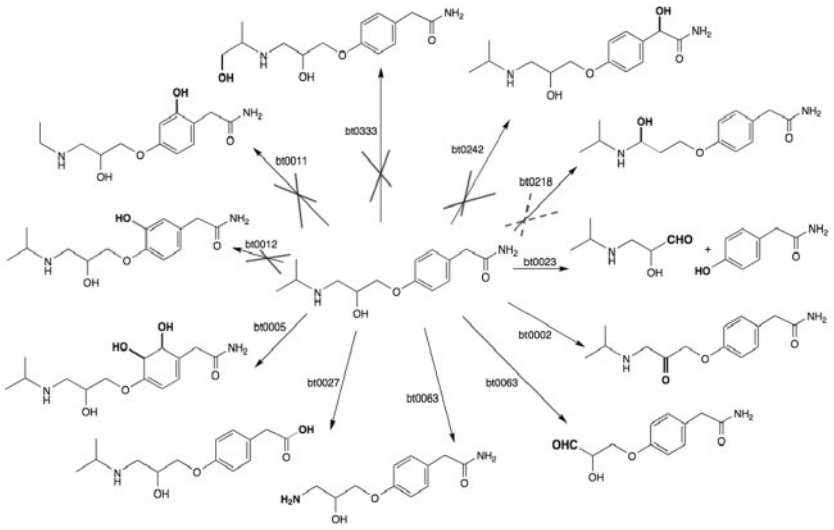

Fig. 2. First generation biotransformation products of atenolol before (all arrows) and after implementation of the extended set of relative reasoning rules (eliminated biotransformations with solid Xs). The biotransformation with a dotted X can be eliminated by absolute reasoning (see text). The transformed atoms and newly formed bonds are marked in bold.

with no relative reasoning associated could still be triggered for the same compound. In contrast to the case for immediate btrules, the products of those rules would also be shown besides the product of the rule that is given priority through an associated relative reasoning rule.

3.3 Validation

Table 4 presents the reduction in predicted transformation reactions achieved by implementing relative reasoning for the two validation sets. Reduction is around 25% for the UM-BBD compound set, and between 10–15% for the xenobiotics. This is as expected, since the relative reasoning rules were trained on the UM-BBD compounds and should therefore work most effectively for them.

More specifically, the UM-BBD compound set includes products of common intermediary metabolism and many of the relative reasoning rules involve btrules only applicable in such common metabolism. This can also be seen when comparing the reductions achieved with and without implementing the immediate rules. Whereas the reduction increases by another 4% for the UM-BBD compounds upon implementing the immediate rules, the xenobiotics are not affected by the immediate rules because these do not apply to the stable parent compounds in the xenobiotics validation set.

In individual cases, the reduction can be considerable as is illustrated in Figure 2 for the human pharmaceutical atenolol. For this compound, the final system including relative and absolute reasoning (aerobic likelihood set to neutral or above) yields a reduction of 42% in the first generation. Still, an average reduction of 16% (xenobiotics) to 27% (UM-BBD) after implementation of the extended set of relative reasoning rules might not at first seem like a large gain. However, if similar reductions are achieved over several generations, this will reduce the number of predicted reactions in a multiplicative way. Statistically, a reduction of 27% in the first generation propagates into a reduction of 47% in the second generation, and a reduction of 61% in the third generation. For the example of atenolol again, final system performance including relative and absolute reasoning over several generations is illustrated in Table 5. The final system shows a reduction of 73% in the third generation of atenolol transformation products, confirming its ability to reduce the hypothetical products to a much more manageable number.

Table 4. Reduction in number of predicted reactions (*PR*) for the UM-BBD and xenobiotics validation sets

		PR (no immediate rules)			Reduction (%) (no immediate rules)	
	n	Original	Strict	Extended	Strict	Extended
UM-BBD	50	429	331 (348)	315 (332)	22.8 (18.9)	26.6 (22.6)
Xenobiotics	47	564	512 (512)	472 (472)	9.2 (9.2)	16.3 (16.3)

Numbers in parentheses ignore immediate rules. Results for before ('original') and after the implementation of both the strict and extended relative reasoning rule set are given.

Table 6 shows the results for the sensitivity and selectivity of UM-PPS before and after implementation of the strict and extended set of relative reasoning rules. For the UM-BBD validation set, about 75% of reported reactions in the database are also predicted by UM-PPS. The sensitivity is lower than 100% because some transformation reactions of UM-BBD compounds are not covered by UM-PPS rules. These include some anaerobic and fungal reactions with only one or two examples in the database as well as multistep reactions where the intermediary products of the pathway are not known. UM-PPS does not assign rules to reactions covering multiple, unknown steps.

The sensitivity for the pesticides validation set is at 74% of known reactions and products predicted, similarly high as for the UM-BBD validation set. We consider this sensitivity to be satisfactory, given the fact that most of the pesticides exhibit considerably more complex structures than the compounds in the UM-BBD. Also, rules for some of the functional groups present in pesticides or biocides

Table 5. Number of predicted transformation products (*PP*) for the β-blocker atenolol over three generations without any rule priorities and with absolute and relative reasoning in place

	PP		Reduction (%)
	No rule priorities	Absolute and relative reasoning	
1st generation	12	7	41.7
2nd generation	80	28	65.0
3rd generation	474	128	73.0

Absolute reasoning criterion set to select products with aerobic likelihood of neutral or above.

Table 6. Sensitivity and selectivity before and after strict and extended relative reasoning for UM-BBD validation set and pesticides/biocides from the xenobiotics validation set

	n	*PR*	KR_p	KR_{np}	Sensitivity (%)	Selectivity (%)
Original						
UM-BBD	50	429	50	17	74.6	11.7
Pesticides	25	280	43	15	74.1	15.4
Strict						
UM-BBD	50	331	50	17	74.6	15.1
Pesticides	25	258	43	15	74.1	16.7
Extended						
UM-BBD	50	315	50	17	74.6	15.9
Pesticides	25	240	43	15	74.1	17.9

PR, number of predicted reactions; KR_p, number of known reactions that are correctly predicted; KR_{np}, number of known reactions that are not predicted.

are currently missing in the UM-PPS. The analysis of the validation results helped us to identify these functional groups. Among others, these include imine bonds and the carbamate functional group, both of which are potentially amenable to enzyme-mediated hydrolytic cleavage.

Sensitivity is identical before and after introduction of relative reasoning. No deterioration of sensitivity is observed between the strict and extended set of relative reasoning rules. This confirms the accuracy of our relative reasoning rules and suggests that the extended set of relative reasoning rules is not overly broad. Due to this, and since the gain in prediction reduction is considerable between the strict and extended set, the extended set was implemented in the current version of the UM-PPS.

In contrast, selectivity is rather low, in the range of 15–18% both for the UM-BBD and pesticide validation sets. Also, it shows only a moderate improvement of 3–4% after the introduction of relative reasoning. This means that even with relative reasoning in place, UM-PPS still has a tendency to predict a considerable number of other transformation products in addition to those that are observed experimentally in the laboratory or under real world conditions.

Although not completely comparable due to slight differences in definitions, the sensitivities of our approach compare well with the sensitivity of the latest version of CATABOL (Dimitrov *et al.*, 2007), which is reported as 70%. However, CATABOL clearly predicts relatively less false positives as reflected in a reported selectivity of 70%. In CATABOL, this selectivity is achieved by restricting the predictions to the application of only the most probable rule from the second generation onwards; the first generation is predicted by applying all possible rules, independent of their probability.

In this context, it is interesting to observe that indeed usually only one product is found for each compound in the UM-BBD, whereas up to five first generation transformation products (average: 2.3 products/compound) are reported for the pesticides. This illustrates an important point: not all possible products that might be formed for a given compound under varying experimental or environmental conditions are usually reported in the UM-BBD, or in the scientific literature on which it is based. This is especially true for pathways identified in pure culture studies, where compounds often represent the sole nutrient and/or energy source and microbes will tend to express the most thermodynamically efficient pathways. Under environmental conditions, on the other hand, co-metabolism is more likely to occur and the product spectrum, therefore, depends more on the enzyme pool available under a given condition rather than on thermodynamic optimization. This situation would typically lead to a broader spectrum of products, which is reflected in the higher number of known products for pesticides compared to UM-BBD compounds. Therefore, the application of only the most probable rule from the second generation onwards in CATABOL seems overly restrictive, especially since it is not detailed whether this procedure was in any way optimized based on existing data. On the other hand, while 15–18% selectivity achieved with relative reasoning in UM-PPS might still be somewhat low, it is probably not far from the optimum selectivity, which lies well below 100%.

4 CONCLUSIONS

This study successfully implemented relative reasoning in biodegradation pathway prediction. The set of relative reasoning rules identified so far effectively reduces the number of predicted transformation products, especially for predictions over several generations of transformation products.

Whereas our results also show that the thus modified UM-PPS system is generally successful at predicting most known products, it still predicts on average five times more products than are experimentally observed. However, this seemingly low sensitivity should be examined in light of the main purpose of the UM-PPS system, which is to support microbiologists, analytical chemists and chemical risk assessors to generate plausible hypotheses regarding possible biotransformation products. It therefore should not be too restrictive.

Atenolol nicely underscores this principle. It was shown to be partially biodegraded during sewage treatment, but no information was given on possible degradation products (Maurer *et al.*, 2007). Figure 2 shows that relative and absolute reasoning leads to a reduction in the first generation biotransformation products of atenolol to seven plausible products. One of them has recently been shown to be the major product in an OECD 308 sediment-water biodegradation study (T.Ternes, personal communication). Screening of a water sample for seven possible transformation products is a feasible task using modern mass spectrometric techniques and would have supported effective discovery of that product.

The set of relative reasoning rules established in this study will be checked for consistency and updated on a yearly basis as more known compounds, reactions and pathways are included into UM-BBD. Also, work on extending the collection of biotransformation rules, and improving the sensitivity of UM-PPS, will continue and focus on some of the missing rules for typical xenobiotic structures identified in this study. Finally, not all the information stored in the rule priority matrix has been uncovered. We are currently exploring machine learning methods to develop chemical structure-based probabilistic relative reasoning rules that would make use of the information present for rule pairs not exhibiting a clear one-directional relationship.

ACKNOWLEDGEMENTS

We thank Chunhui Li for the initial code to produce the rule priority matrix; Michael Turnbull and Rachael Long for writing and revising many of the UM-PPS btrules; and John Bumpus and Jack Richman for critical review of the article.

Funding: Fellowship was granted for advanced researchers from the Swiss National Science Foundation (PA002-113140 to K.F.), Lhasa Limited, the US National Science Foundation (NSF0543416), and the University of Minnesota Supercomputing Institute.

Conflict of Interest: none declared.

REFERENCES

Battaglin,W.A. *et al.* (2005) Glyphosate, other herbicides, and transformation products in Midwestern streams, 2002. *J. Am. Water Resour. Assoc.*, **41**, 323–332.

Button,W.G. *et al.* (2003) Using absolute and relative reasoning in the prediction of the potential metabolism of xenobiotics. *J. Chem. Inf. Comp. Sci.*, **43**, 1371–1337.

Dimitrov,S. *et al.* (2007) A kinetic model for predicting biodegradation. *SAR QSAR Environ. Res.*, **18**, 443–457.

Ellis,L.B.M. *et al*. (2006) The University of Minnesota biocatalysis/biodegradation database: the first decade. *Nucleic Acids Res.*, **34**, D517–D521.

Embrechts,M.J. and Ekins,S. (2007). Classification of metabolites with kernel-partial least squares (K-PLS). *Drug Metab. Dispos*., **35**, 325–327.

EMEA (2006) *Guideline on the Environmental Risk Assessment of Medicinal Products for Human Use. Committee for Medicinal Products of Human Use (CHMP).* European Medicines Agency (EMEA), London.

Gomez,M.J. *et al.* (2007) The environmental fate of organic pollutants through global microbial metabolism. *Mol. Syst. Biol.*, **3**, 114.

Hou,B.K. *et al.* (2004) Encoding metabolic logic: predicting biodegradation. *J. Ind. Microbiol. Biotechnol.*, **70**, 261–272.

Jaworska,J. *et al.* (2002) Probabilistic assessment of biodegradability based on metabolic pathways: catabol system. *SAR QSAR Environ. Res.*, **13**, 307–323.

Jaworska,J. *et al.* (2003) Recent developments in broadly applicable structure-biodegradability relationships. *Environ. Toxicol. Chem.*, **22**, 1710–1723.

Klopman,G. and Tu,M.H. (1997) Structure-biodegradability study and computer-automated prediction of aerobic biodegradation of chemicals. *Environ. Toxicol. Chem.*, **16**, 1829–1835.

Kolpin,D. *et al.* (2004) Degradates provide insight to spatial and temporal trends of herbicides in ground water. *Ground Water*, **42**, 601–608.

Maurer,M. *et al.* (2007) Elimination of β-blockers in sewage treatment plants. *Water Res.*, **41**, 1614–1622.

Mu,F. *et al.* (2006) Prediction of oxidoreductase-catalyzed reactions based on atomic properties of metabolites. *Bioinformatics*, **22**, 3082–3088.

Oh,M. *et al.* (2007) Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways. *J. Chem. Inf. Model*, **47**, 1702–1712.

REACH (2006) Regulation (EC) No 1907/2006 of the European Parliament and of the Council of 18 December 2006 concerning the registration, evaluation, authorisation and restriction of chemicals. *Official J. Eur. Union*, **49**, L396.

Roberts,T.R., (ed.) (1998) *Herbicides and plant growth regulators.* Vol. Part 1. The Royal Society of Chemistry, Cambridge.

Urbanczik,R. and Wagner,C. (2005) Functional stiochiometric analysis of metabolic networks. *Bioinformatics*, **21**, 4176–4180.

VICH (2004) *Environmental Impact Assessment (EIAS) for Veterinary Medicinal Products – Phase II Guidance*. International Cooperation on Harmonisation of Technical Requirements for Registration of Veterinary Medicinal Products (VICH), Brussels.