Machine Learning and Data Mining Approaches to Biodegradation Pathway Prediction

Jörg Wicker[1], Kathrin Fenner[2,3], Lynda Ellis[4], Larry Wackett[5], and Stefan Kramer[1]

[1] TU München, Institut für Informatik/I12,
D-85748 Garching b. München, Germany
[2]Swiss Federal Institute for Aquatic Science and Technology (Eawag),
CH-8600 Dübendorf, Switzerland
[3]Institute of Biogeochemistry and Pollutant Dynamics, Swiss Federal Institute of Technology (ETH), CH-8092 Zürich, Switzerland
[4]Department of Laboratory Medicine and Pathology,
University of Minnesota, Minneapolis, MN 55455, USA
[5]Department of Biochemistry, Molecular Biology and Biophysics,
University of Minnesota, St. Paul, MN 55108, USA
joerg.wicker@in.tum.de, kathrin.fenner@eawag.ch, ellis@msi.umn.edu,
wacke003@umn.edu, stefan.kramer@in.tum.de

1 Background

Modeling biological processes is among the central research goals of systems biology. One of the most prominent and challenging problems in this area is the prediction of chemical reactions and pathways (i.e., chains of reactions). We are tackling the problem of pathway prediction in the context of biodegradation. Current methods for the prediction of reaction products and pathways either employ knowledge-based [1] or machine learning based [2] approaches. In this paper, we investigate ways to combine the two approaches, where we assume a given set of biotransformation rules and learn so-called *relative reasoning rules* [3], both logical and probabilistic, to control the application of rules. The methods are developed for the University of Minnesota Biocatalysis/Biodegradation Database (UM-BBD) [4], and the rules from its pathway prediction system (UM-PPS) [5].

2 UM-BBD and UM-PPS

For over a decade, the University of Minnesota Biocatalysis/Biodegradation Database (UM-BBD, http://umbbd.msi.umn.edu/) has compiled, stored and displayed data on microbial catabolism of environmental pollutants on the web. The scope of the UM-BBD pathway collection emphasizes the wide range of different functional groups present in organic molecules and the corresponding great breadth of microbial metabolic reactions that are described in the scientific literature. Currently, biotransformation schemes for mostly xenobiotic compounds are accessible through UM-BBD, encompassing roughly 1,000 parent compounds and intermediates.

In 2002, the University of Minnesota Pathway Prediction System (UM-PPS, http://umbbd.msi.umn.edu/predict/) was built based on the collection of known reactions and pathways in the UM-BBD. The UM-PPS accepts user compounds and proposes plausible microbial catabolic reactions and pathways for compounds for which no experimental data exists. The system is built on substructure search, atom-to-atom mapping, and over 200 biotransformation rules, which cover over 90% of appropriate reactions in the UM-BBD. UM-PPS biotransformation rules are generalizations and abstractions of known reactions from the database. If a rule matches certain functional groups on its left-hand side, then it transforms the structure according to its right-hand side.

As any rule-based system that predicts sequences of transformation steps, UM-PPS suffers from combinatorial explosion. For an illustration of this, see the first two generations of degradation products predicted by UM-PPS in Figure 1. The problem is particularly aggravated for the structurally more complex contaminants of current concern, e.g., pesticides, biocides or pharmaceuticals. Potential users of the system such as environmental microbiologists, risk assessors, and analytical chemists are overwhelmed by the number of possible products, and find it hard to identify the most plausible products. In an initial effort to restrict combinatorial explosion, expert knowledge was used to rank the rules in UM-PPS into five aerobic likelihood groups (very likely, likely, neutral, unlikely, and very unlikely), which can be used to reduce the number of predictions for aerobic conditions, e.g., by removing unlikely and very unlikely biotransformations. However, only 20% of the UM-PPS rules have aerobic likelihoods of unlikely or very unlikely, which makes this approach alone insufficient.

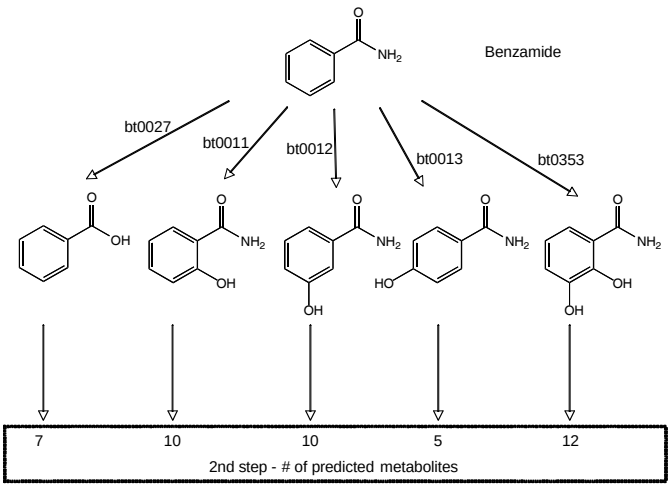

Fig. 1. Two generations of metabolites predicted by UM-PPS

3 Data-Driven Extraction of Relative Reasoning Rules

In a first approach, we automatically extracted relative reasoning rules [3] to limit combinatorial explosion [6]. The concept of relative reasoning is to give priority to those of a set of applicable rules that encode the more plausible biotransformations. Thereby the information on such rule priorities is extracted from known biodegradation pathways, in our case the UM-BBD.

In short, known biotransformation reactions for roughly 700 compounds contained in UM-BBD were compared to all possible, aerobic UM-PPS biotransformation rules (btrules) triggered by these compounds. The results were stored in a data matrix, distinguishing between three cases: Rules that are not triggered by a given compounds, rules that are triggered but do not represent a known biotransformation, and finally rules that are triggered and do represent a known biotransformation. Based on the training matrix, pairs of rules were sought that show a clear one-directional relationship, i.e., where the known transformations for all compounds that trigger both rules proceed exclusively according to one of the two rules. 109 relative reasoning rules were identified that fulfilled this criterion for a sufficiently large and diverse set of compounds. The thus obtained rule priorities were implemented in two ways: (i) as relative reasoning rules, i.e., whenever two rules for which a relative reasoning rule exists are triggered for a given compounds, only the product of the rule with higher priority is shown, or (ii) as immediate rule in the case of rules that had a higher priority in all relative reasoning rules. Whenever an immediate rule is applicable, only the product of the immediate rule is shown to the user and the user is not given the choice of selecting any of the other theoretically possible transformation products.

The final system was tested against an internal (UM-BBD compounds) and an external validation set (diverse set of 50 pesticides, biocides and pharmaceuticals). A reduction of about 25% in the number of predicted first generation transformation products was achieved upon implementation of relative reasoning, while the percentage of correctly predicted, experimentally known products remained at 75% (recall). While the precision improved by 3-4% upon introduction of relative reasoning, it remained in the range of 0.15 to 0.20.

4 Machine-Learning Based Relative Reasoning

In summary, the relative reasoning approach was very easily implemented into the working UM-PPS and due to the restrictiveness with which relative reasoning rules were selected, i.e., no exception tolerated, the recall of the system remains untouched. However, due to the very same restrictivness, also in applying the relative reasoning rules, i.e., only yes/no answers regarding the applicability of a rule are possible, precision could only be improved marginally. Therefore, we developed a more statistically oriented approach to improve the system's precision based on the same set of information that was used to derive relative reasoning rules [7]. More precisely, we address the following problem: Given the structure of a substrate and the rules triggering for that substrate, which of the transformations suggested by the rules should be accepted? Our proposed solution is based on one classifier per rule. The

input of a classifier is the chemical structure of the substrate and the set of alternative transformation rules triggering for that substrate. The output is the probability that the transformation product suggested by the rule is actually observed. Clearly, the decision to accept a product or not can be made dependent on this probability: The application of individual rules can be tuned such that only transformations above a certain probability threshold are accepted. In this way, its also possible to control the generality of whole rule sets and the overall number of products. Thus, it is easy to address the fundamental trade-off between the completeness and the accuracy of predictions. In technical terms, we can analyze the performance of both individual rules and systems in recall-precision space, and visualize their performance in two-dimensional plots. Moreover, it is possible to explicitly choose a suitable point in recall-precision space. If all rules are used for prediction, the proposed method achieves a recall R=0.72 and a precision P=0.60 for Random Forests [8] in ten-fold cross-validation over the UM-BBD compounds, and, at another point in recall-precision space, R=0.68 and P=0.65. If only classifiers with a sufficient number of examples and a balanced class distribution are selected (here, a subset of 13 rules), it is possible to improve this result to R=0.8 and P=0.8.

References

1. Greene, N. et al.: Knowledge-based Expert Systems for Toxicity and Metabolism Prediction: DEREK, StAR, and METEOR. SAR &QSAR in Environmental Research, 1999, 10, 299-313 (1999)
2. Jaworska, J. et al.: Probabilistic Assessment of Biodegradability Based on Metabolic Pathways: CATABOL system. SAR &QSAR in Environmental Research, 13, 307-323 (2002)
3. Button, W.G. et al.: Using Absolute and Relative Reasoning in the Prediction of the Potential Metabolism of Xenobiotics. J. Chem. Info. Comp. Sci., 43, 1371-1377 (2003)
4. Ellis, L.B.M. et al.: The University of Minnesota Biocatalysis/Biodegradation Database: the First Decade. Nucleic Acids Res., 34, D517-D521 (2006)
5. Hou, B.K. et al.: Encoding Metabolic Logic: Predicting Biodegradation. J. Ind. Microbiol. Biotech., 70, 261-272 (2004)
6. Fenner, K. et al.: Data-Driven Extraction of Relative Reasoning Rules to Limit Combinatorial Explosion in Biodegradation Pathway Prediction, Bioinformatics (accepted for publication) (2008)
7. Wicker, J. et al.: Predicting Biodegradation Products and Pathways: A Hybrid Knowledge-Based and Machine Learning Approach, to be submitted (2008)
8. Breiman, L.: Random Forests, Machine Learning, 45, 5-32 (2001)