Symbolic Inference of Xenobiotic Metabolism

D.C. McShan, M. Upadhyaya, and I. Shah

Pacific Symposium on Biocomputing 9:545-556(2004)

SYMBOLIC INFERENCE OF XENOBIOTIC METABOLISM

D.C. MCSHAN, M. UPDADHAYAYA and I. SHAH
School of Medicine
University of Colorado
4200 East 9th Avenue, B-119
Denver, CO 80262
{daniel.mcshan,minesh.upadhyaya,imran.shah}@uchsc.edu

Abstract

We present a new symbolic computational approach to elucidate the biochemical networks of living systems *de novo* and we apply it to an important biomedical problem: xenobiotic metabolism. A crucial issue in analyzing and modeling a living organism is understanding its biochemical network *beyond* what is already known. Our objective is to use the available metabolic information in a representational framework that enables the inference of novel biochemical knowledge and whose results can be validated experimentally. We describe a symbolic computational approach consisting of two parts. First, biotransformation rules are inferred from the molecular graphs of compounds in enzyme-catalyzed reactions. Second, these rules are recursively applied to different compounds to generate novel metabolic networks, containing new biotransformations and new metabolites. Using data for 456 generic reactions and 825 generic compounds from KEGG we were able to extract 110 biotransformation rules, which generalize a subset of known biocatalytic functions. We tested our approach by applying these rules to ethanol, a common substance of abuse and to furfuryl alcohol, a xenobiotic organic solvent, which is absent in metabolic databases. In both cases our predictions on the fate of ethanol and furfuryl alcohol are consistent with the literature on the metabolism of these compounds.

Introduction

The objective of this work is to develop a predictive strategy for elucidating metabolism. We mold available metabolic information in an expressive symbolic representation and employ a novel inference framework to explore unchartered pathways. We hypothesize that biochemical rules can be inferred from the databases of endogenous metabolism and that we can use these rules to predict the metabolism of unknown xenobiotics through detoxification pathways. In particular, we focus on xenobiotic pathways in mammalian systems.

What is the importance of discovering new pathways? Our knowledge of metabolism is essentially incomplete and it can be argued that cataloging

all possible mammalian xenobiotic pathways is infeasible. With the availability of the complete genomic blueprint for living systems and a large set of known biotransformations, it is becoming possible to theoretically elucidate metabolism. This includes the analysis of endogenous as well as xenobiotic pathways. Drugs, substances of abuse and environmental pollutants are examples of compounds that may not occur naturally in a living system. Since these compounds and/or their metabolic by-products can be potentially toxic, investigating xenobiotic metabolism is important for human health and the environment.

Pathway inference is a computationally challenging problem even with the availability of the genomic blueprint for a living system and the functional annotations of its putative genes. Since the availability of the first microbial genome, *Haemophilus influenza*,[2] a number of metabolic reconstruction tools have been developed. These include PathoLogic,[7] MAGPIE,[3,4] and PathFinder[5]. These methods focused on matching putatively identified enzymes with known, or "reference", pathways. Although reconstruction is an important starting point for metabolic processes it does not enable the discovery of new pathways. To overcome some of these issues we have recently developed a new pathway inference system to search for novel metabolic routes called PathMiner[7]. PathMiner uses known biotransformations to synthesize new pathways and employs heuristics to contain the combinatorial complexity of the search. This paper delves into a deeper biological problem: *de novo* pathway inference and its practical application to a biomedical problem: xenobiotic metabolism.

The metabolic potential of a living system depends on biocatalysis. However, understanding the mechanisms of enzymatic catalysis is an extremely difficult problem, and knowledge in this area is limited to a handful of well-studied examples. Generally, biochemists can abstract empirical "rules" for the biotransformation of metabolites by enzymes. For instance, consider the broad range of substrates for *Saccharomyces cerevisae* (yeast) alcohol dehydrogenase (YADH), which reduces acetaldehyde and a variety of other aldehydes[1], and oxidizes ethanol, and other acyclic primary alcohols. Yet an alcohol dehydrogenase from *Thermoanaerobium brokii* (TADH) catalyzes the stereospecific reduction of ketones and the oxidation of secondary alcohols. The functions of YADH and TADH share common attributes and have some unique differences: they are both alcohol dehydrogenases but their specificities for the alcohols are different. The functions of these enzymes can be expressed in terms of the functional groups modified (alcohol to aldehyde or ketone), and the backbone structure of the molecule (primary or secondary alcohol). This is essentially a symbolic description of biocatalysis and we believe that it can be applied to

complete metabolic systems.

Methods

Our strategy for elucidating *de novo* xenobiotic metabolism consists of two main steps. First, we use biotransformation data to derive symbolic chemical substructural rules that generalize the action of enzymes on specific compounds. Second, we apply these rules iteratively to a compound to generate a plausible metabolic system. We describe these steps in the following sections but first we discuss our metabolic representation.

Representing biotransformations and rules

Our abstraction of metabolic concepts is based on work by Karp[7] in terms of the high-level concepts including pathways, enzyme-catalyzed reactions and transformations. At the level of biotransformations we are motivated by Kazic[9] in that we focus on the specific chemical substructural details of metabolites that are modified through biocatalysis. In our system, compounds are represented as X. Compounds in our abstraction have a chemical structure which is represented as a molecular graph, Γ, in which nodes are atoms and edges are bonds. In the context of a biotransformation the pattern of substructural changes from the input compound to the output compound is represented as a rule, U. A rule captures the concept of functional group changes that occur in a biotransformation. Rules are implicitly unidirectional so reversible transformations are represented as two separate rules. The two molecular graphs of a rule are indicated by the input graph, Δ^-, and the output graph, Δ^+. For instance, the rule for the conversion of a primary alcohol to an aldehyde is shown in Figure 1. In this case Δ^- is an alcohol moiety, which is converted to Δ^+, an aldehyde moiety.

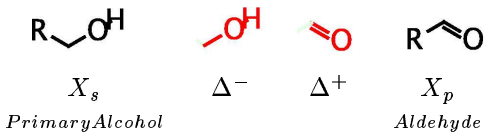

Figure 1: Alcohol dehydrogenase (EC 1.1.99.9) Transformation from abstract PrimaryAlcohol to abstract Aldehyde showing the computedΔ^- and Δ^+ moieties. The Δ^-moeity is the subgraph that is in $X_S$but not in X_p. The Δ^+moeity is the subgraph that is in $X_p$but not in X_s.

In the present work we focus on changes at the level of functional groups

between pairs of compounds. We represent the conversion of one input compound to one output compound as a transformation. This simplifies our representation of reactions in terms of the main metabolites. In this work we obtain this data from the KEGG distribution, but we are also exploring automated methods for identifying the main metabolites in a reaction.

Extracting transformation rules from reaction data

One strategy for identifying rules is to curate them manually, however, our goal is to use the available metabolic data[8,7] to derive biotransformation rules automatically. This is a difficult problem in general as the information about reactive moieties is not explicitly available. In this paper we have used a simple strategy for extracting rules automatically from "general" reactions. In KEGG, for instance, general reactions are defined when the input and the output compounds are both Markush structures. We find 741 general reactions in KEGG, which constitute 20% of the reactions annotated as being human-specific. For example a gene that is extremely important in xenobiotic metabolism and encodes cytochrome P-450 enzyme, CYP2D6, is implicated in the disposition of over thirty toxins. In KEGG, the P-450 enzyme (EC 1.14.14.1) is associated with only four reactions as shown in Figure 2. There are two specific reactions involved in endogenous functions associated with tryptophan metabolism and gamma-hexachlorocyclohexane degradation. The other two operate on general compounds denoted by their Markush structures (these are abstract structures containing a wildcard "R" group and specific functional groups). We convert these general reactions automatically to rules as described above. This is done by replacing the wildcard of the substrate with "C" and storing it as the Δ^- subgraph in the resulting rule; similarly, the "R" in the product graph is replaced and the resulting graph is stored as Δ^+. To our knowledge no one has taken advantage of this annotation before in metabolic pathway inference.

In this work we focus on the rules important in xenobiotic metabolism in mammalian systems, including oxidation, reduction, hydrolysis and conjugation to mention a few. There are generally two phases in xenobiotic metabolism. In phase 1 the compounds are 'functionalized', which means that a reactive functional group is exposed. Detoxification occurs in phase 2 by further action on the functional groups, which is the form in which the compound is excreted. For instance, the first phase activates a molecular oxygen in the input compound, and the second phase conjugates it. Glucuronidation is the most common conjugate and can be attached to any labile oxygen. In the case of alcohol metabolism, both the alcohol and the acid can usually be conjugated.

$$
\begin{aligned}
\textit{Melatonin} &\Leftrightarrow 6-\textit{Hydroxymelatonin} \\
\textbf{FattyAcid} &\Leftrightarrow \textbf{alphaHydroxyFattyAcid} \\
\textbf{Alkane} &\Leftrightarrow \textbf{ROH} \\
\textit{Parathion} &\Leftrightarrow \textit{Paraoxon}
\end{aligned}
$$

Figure 2: CYP2D6 (EC 1.14.14.1) reactions in KEGG. Compounds are either **Abstract**(contain one or more Markush "R" groups) or *Normal* (have unique structure).

Biotransformation rule application

Our rule application algorithm is illustrated in Algorithm 1. A rule is applied to a substrate X_s by searching the graph of X_s, Γ_s for the subgraph Δ^-. If the subgraph Δ^- is found, it is replaced by the Δ^+ graph to yield the product graph, Γ_p. This is summarized as follows:

$$\Gamma_s - \Delta^- + \Delta^+ \Rightarrow \Gamma_p.$$

This is graphically illustrated in Figure 3.

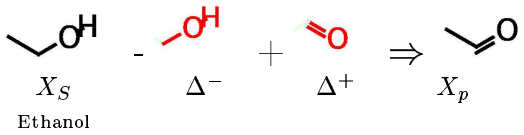

Figure 3: Application of alcohol dehydrogenase rule to ethanol

The product of applying a rule to a compound can be a completely novel compound or a known compound. We use subgraph isomorphism to search the product molecular graph against the database of known compounds. If the compound is not found, a novel compound $X_p^{'}$ is created and given a unique identifier (Nxxxxxx in which x is a digit from 0-9). The corpus of all rules is designated $\overline{U}$. We have a top-level function `metabolize`(X,$\overline{U}$,n) which takes a compound X and systematically applies each rule in the rule-base $\overline{U}$ through n iterations.

```
input  : X_s, compound to metabolize
         Ū, list of rules
         n, iterations
output : Graphical visualization
         Products
Products ← φ
Γ_s ← molecular-graph(X_s)
for (Δ⁻, Δ⁺) ← Ū do
    Γ_p ← graph-replace(Γ_s,Δ⁻,Δ⁺)
    if Γ_p then
    X_p ← find-compound-by-graph(Γ_p)
    if X_p = φ then X_p ← make-novel-compound(Γ_p)
    pushnew(X_p,Products)
if n > 1 then
    for X in Products do
        append(metabolize(X,Ū,n − 1),Products)
```

Algorithm 1: `metabolize`(X,$\bar{U}$,n). Algorithm to create a network of pathways length n from input compound X_s by applying rules $\bar{U}$. Initially the list of *Products* is set to null. The molecular graph, Γ_s, of the input compound is obtained from the KEGG `mol` file representation. For every rule in the rulebase $\bar{U}$, we obtain the Δ^- and Δ^+ subgraphs. The product graph, Γ_p is obtained by performing a graphical search/replace on the input graph, Γ_s. If Γ_p is non, i.e., a match was found and applied, then the product graph Γ_p is searched against the database of known compounds and the database of novel compounds to see if an isomorphic graph exists. If the graph matches an existing compound, then X_p is returned. If there is no identified compound with the graph, then a novel compound, X_p is generated and given a unique identifier (the Nxxxxx symbols in the diagrams). In either case, the product, X_p is pushed onto the *Products* list for this metabolite X_s. This process can occur iteratively for every product, X in the *Products* list. The `metabolize` function is simply called again with the recursion level reduced. The results are appended to the *Products* list.

Implementation

The system is implemented in Allegro Common Lisp. The metabolic databases are read in and parsed into CLOS structures. For visualization, the transformations are exported to the AT&T graphviz program neato which does a simple force-based layout of the metabolic graph. This network is read back in and presented with the nodes replaced by compound structures using our internal visualization system. The novel compounds that are produced by the appli-

U_x	Reactant	Product	E.C.	Enzyme
1	ALCOHOL ROH	β-L-ARABINOSIDE $RC_5H_9O_5$	3.2.1.88	VICIANOSIDASE
2	ALKYL SULFATE RO_4HS	ALCOHOL ROH	2.3.1.84	ALCOHOL ACETYLTRANSFERASE
3	ALCOHOL ROH	ACETYL ESTER $RC_2H_3O_2$	2.3.1.84	ALCOHOL ACETYLTRANSFERASE
4	ALCOHOL ROH	GLUCURONIDE $RC_6H_9O_7$	3.2.1.31	KETODASE
5	FATTY ACID $RCHO_2$	α-OH FATTY ACID $RC_3H_5O_3$	1.14.14.1	MICROSOMAL P-450
6	R-CN RCN	MCA AMIDE RCH_2NO	4.2.1.84	NHASE
7	1-ALCOHOL RCH_3O	ALDEHYDE $RCHO$	1.1.99.20	ALKAN-1-OL DEHYDROGENASE
8	ALDEHYDE $RCHO$	FATTY ACID $RCHO_2$	1.2.99.3	ALDEHYDE DEHYDROGENASE
9	ALDEHYDE $RCHO$	R-COOH RCO_2	1.2.3.1	ALDEHYDE OXIDASE
10	R-COOH RCO_2	ALDEHYDE $RCHO$	1.2.3.1	ALDEHYDE OXIDASE

Table 1: Simplest 10 of 110 rules inferred from KEGG generic reactions

cation of the rules are simply graphs. In order to visualize the compounds, we require 2D coordinates. To achieve this, we export the graph as a mol file with the 2D coordinates as zeroes and then layout the mol file using the JChem molconvert package. The mol files are read back in and stored with the compounds as they are created.

Results and Discussion

We used a recent version of the KEGG database which had 10,635 compounds, out of which 825 are generic. Of the 5,428 reactions in the KEGG database, 741 operate on the generic compounds. From this data, we infer 110 biotransformation rules, and the 10 simplest ones are summarized in Table 1. These rules correspond to enzymes which have flexibility in the substrates they can transform.

Using our symbolic computational approach described in the previous sections we elucidate the *de novo* metabolism of two compounds. First, we consider ethanol, which is a common substance of abuse and for which we have some data of human metabolism.

Second, we demonstrate the fate of furfuryl alcohol, which is is an industrial organic solvent used as a paint thinner and is absent in our database. Experimental evidence suggests that prolonged exposure to furfuryl alcohol may have significant toxicological effects. We first apply the rules to the compound ethanol which is in the database. The graph is shown in Figure 4. Next, we apply the rules to a new compound, furfuryl alcohol, which is not in the database. The result is shown in Figure 5. That some of the

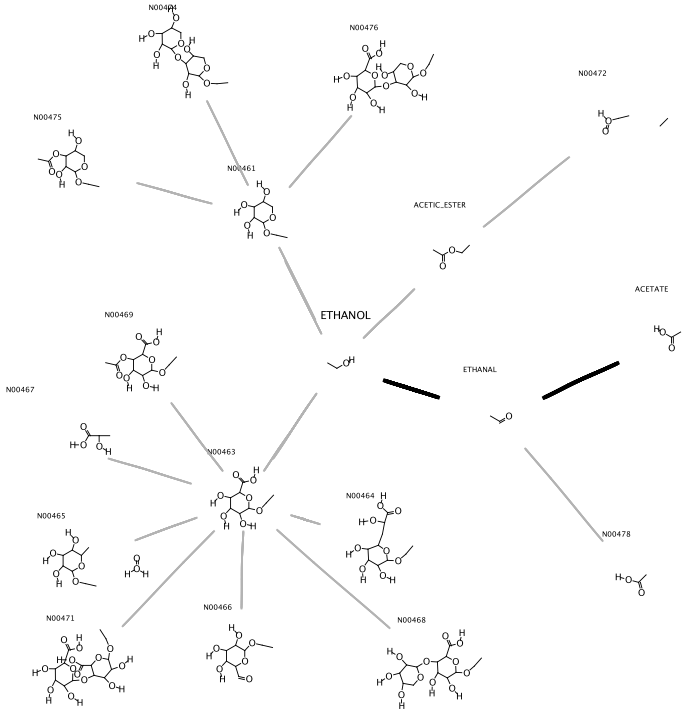

Figure 4: The *de novo* prediction of ethanol metabolism. Ethanol is in the center of the figure. The highlighted transformations are the activation of the alcohol to an aldehyde by alcohol dehydrogenase (EC 1.1.99.20), then to an acid by aldehyde oxidase (EC 1.2.3.1), respectively. Not shown, but in the next iteration is the O-glycosylation of the aldehyde by beta-Glucuronidase (EC 3.2.1.31).

nodes in our ethanol metabolism graph match to known compounds in the database is encouraging. Additionally, we were able to identify the pathway, *alcohol* $\Rightarrow$ *aldehyde* $\Rightarrow$ *acid* $\Rightarrow$ *conjugation*, which recapitulates the standard ethanol detoxification pathway. We are also able to predict metabolites for a compound previously unknown to the system. The furfuryl alcohol metabolic predictions are consistent with literature. Martin, *et. al.*, report that furfuryl alcohol can be O-glycosylated by beta-Glucuronidase[10] as we predict (shown as compound N00482 in Figure 5). Additionally, the acid of furfurol, 2-furoate, is actually in the KEGG database and is identified as such by the algorithm. Nomeir, et. al., report that the initial step in furfuryl alcohol metabolism in rat is the oxidation to furoic acid, which is excreted unchanged and decarboxy-

lated, or conjugated with glycine or condensed with acetic acid[11]. In this case, the limitations in our system to predict the condensation with acetic acid, for instance, lie in the breadth of the rules, not in the fundamental methodology. By extending our method for inferring new rules based on known biochemistry we can overcome this limitation.

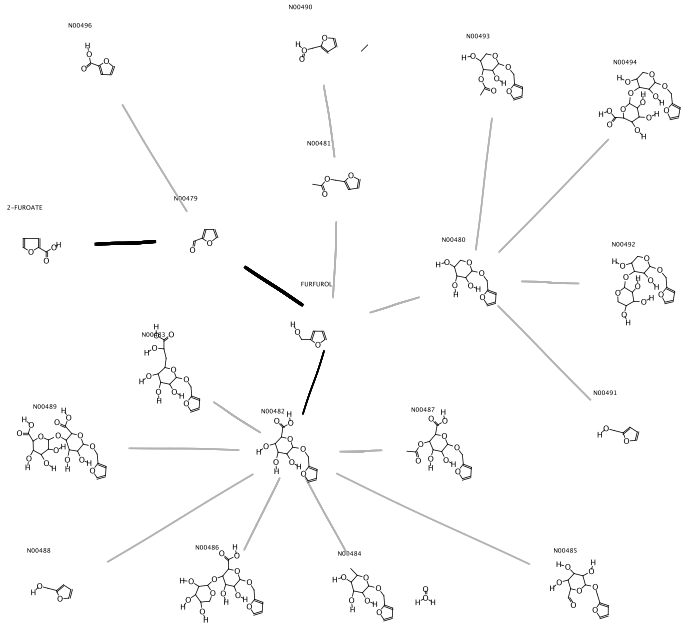

Figure 5: The *de novo* prediction of furfuryl metabolism. Furfurol is in the center of the figure. The highlighted transformation between compound furfurol and compound N00482 (up and to the left) is an O-glycosylation by beta-Glucuronidase (EC 3.2.1.31). The highlighted transformations below furfurol are the activation of the alcohol to an aldehyde (N00479, furfural) by alcohol dehydrogenase (EC 1.1.99.20), then to an acid by aldehyde oxidase (EC 1.2.3.1), respectively. The acid is identified by the algorithm as being in the KEGG database (by graph similarity) as 2-Furoate (C01546). In the next iteration, not shown, the acid is finally O-glycosylated by beta-Glucuronidase (EC 3.2.1.31).

Most of the complex products of furfuryl alcohol are simply consecutive glucurodinations by the rule:

$$Alcohol \quad \Rightarrow \quad \beta - D - Glucuronide$$

Due to the lack of specificity of this rule to primary alcohols, glucuronidation is applied to the hydoxyl groups on the $\beta - D - Glucuronide$. While this might be

biologically valid, in reality, glucuronidation renders a compound water soluble after which it is eliminated by excretion. This limitation is beyond the scope of the current work but can be addressed in the future by considering the physical properties of compounds, like water-solubility.

That a biotransformation rule *can* be applied does not imply that it *is* biochemically valid. For instance, consider the biotransformation rules that apply to a hydroxyl functional group. Compounds containing this functional group include primary alcohols, secondary alcohols, and also carboxylic acids. Enzymes that act on alcohols may not act on carboxylic acids and vice-versa. To capture the substrate specificity of enzymes we are working on a more sophisticated representation of rules that can improve their biological validity. Though this is a limitation of our present algorithm, our predictions are still useful for elucidating potential xenobiotic metabolism, which can be tested experimentally.

It is important to contrast our approach to other rule-based approaches[7,6] in pathway prediction. One of the main advantages of our strategy is automated biotransformation rule extraction from available resources of metabolic data. As opposed to the manual curation-based efforts, our approach will scale gracefully with increasing data for two important reasons. First, our algorithm for rule extraction can be extended to utilize most of the available enzyme-catalyzed reaction data beyond the generic reactions in KEGG. Second, we can control the combinatorial explosion of plausible biotransformations by extending our existing algorithm on pathway search[7]. Another advantage of our approach is that we can relate our biotransformation predictions to the organism-specific enzymes and genes, which is crucial for *in vivo* or *in vitro* experimental validation.

Conclusion

We have developed a symbolic inference approach and demonstrated the *de novo* elucidation of metabolism. This was accomplished by representing biocatalysis, which is the basis of metabolism, in terms of expressive symbolic biotransformation rules. These biotransformation rules generalize the biocatalytic functions of enzymes and enable the discovery of new metabolic potential in living systems. We developed an algorithm to extract these rules from known enzyme-catalyzed reactions and to apply these rules to elucidate the metabolism of new compounds. We successfully tested this concept to predict the xenobiotic metabolism of ethanol and furfuryl alcohol. The results are encouraging because furfuryl alcohol is absent in our database and yet we can correctly identify its products through O-glycosidation and oxidation to

furoic acid in agreement with the literature. These results are also biologically interesting because they support the notion that xenobiotic metabolism is a manifestation of endogenous biocatalytic abilities in an organism. Though there are a some limitations in our approach the method is quite general and scalable for investigating the metabolic network of any living system.

This work supports the relevance of symbolic approaches in discovering the biochemical capabilities of living systems. Our results on xenobiotic metabolism offer a prelude to the potential discoveries that can be made in combination with high-throughput or traditional experimental strategies.

Aknowledgments

The authors acknowledge Weiming Zhang for the visualization software. This work is sponsored by the National Science Foundation (BES-9911447), the Department of Energy (DE-FG03-01ER63111/M003), and the Office of Naval Research (N00014-00-1-0749).

References

1. *Applications of Biochemical Systems in Organic Chemistry.* Wiley, New York, N.Y., 1976.
2. R.D. Fleischmann et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269:469–512, 1995.
3. T. Gaasterland and E.E. Selkov. Automatic Reconstruction of Metabolic Networks Using Incomplete Information. *ISMB*, 3:127–135, 1995.
4. T Gaasterland and CW Sensen. MAGPIE: automated genome interpretation. *Trends Genet*, 12(2):76–78, 1996.
5. A. Goesmann, M. Haubrock, F. Meyer, J. Kalinowski, and R. Giegerich. PathFinder: reconstruction and dynamic visualization of metabolic pathways. *Bioinformatics*, 18(1):124–9, 2002. 11836220.
6. B.K. Hou, L.P. Wackett, and L.B. Ellis. Microbial pathway prediction: a functional group approach. *J Chem Inf Comput Sci*, 43(3):1051–7, 2003.
7. P. Karp and M. Riley. Representations of metabolic knowledge: Pathways. In R. Altman, D. Brutlag, P. Karp, R. Lathrop, and D. Searls, editors, *Second International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, 1994.
8. P.D. Karp, M. Krummenacker, S.M. Paley, and J. Wagg. Integrated pathway/genome databases and their role in drug discovery. *Trends in Biotechnology*, 17(7):275–281, 1999.

9. T Kazic. Reasoning about biochemical compounds and processes. pages 35–49. World Scientific, Singapore, 1992.
10. B.D. Martin, E.R. Welsh, J.C. Mastrangelo, and R. Aggarwal. General O-glycosylation of 2-furfuryl alcohol using beta-glucuronidase. *Biotechnol Bioeng*, 80(2):222–7, 2002.
11. A.A. Nomeir, D.M. Silveira, M.F. McComish, and M. Chadwick. Comparative metabolism and disposition of furfural and furfuryl alcohol in rats. *Drug Metab Dispos*, 20(2):198–204, 1992.