Published online 4 June 2008 Nucleic Acids Research, 2008, Vol. 36, Web Server issue W427–W432
doi:10.1093/nar/gkn315

The University of Minnesota pathway prediction system: predicting metabolic logic

Lynda B.M. Ellis[1,]*, Junfeng Gao[1], Kathrin Fenner[2,3] and Lawrence P. Wackett[4]

[1]Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN, 55455, USA, [2]Institute of Biogeochemistry and Pollutant Dynamics (IBP), ETH Zürich, CH-8092 Zürich, [3]Swiss Federal Institute for Aquatic Science and Technology (Eawag), CH-8600 Dübendorf, Switzerland and [4]Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, St Paul, MN 55108, USA

Received January 24, 2008; Revised April 17, 2008; Accepted May 7, 2008

ABSTRACT

The University of Minnesota pathway prediction system (UM-PPS, http://umbbd.msi.umn.edu/predict/) recognizes functional groups in organic compounds that are potential targets of microbial catabolic reactions, and predicts transformations of these groups based on biotransformation rules. Rules are based on the University of Minnesota biocatalysis/biodegradation database (http://umbbd.msi.umn.edu/) and the scientific literature. As rules were added to the UM-PPS, more of them were triggered at each prediction step. The resulting combinatorial explosion is being addressed in four ways. Biodegradation experts give each rule an *aerobic likelihood* value of Very Likely, Likely, Neutral, Unlikely or Very Unlikely. Users now can choose whether they view all, or only the more aerobically likely, predicted transformations. *Relative reasoning*, allowing triggering of some rules to inhibit triggering of others, was implemented. Rules were initially assigned to individual chemical reactions. In selected cases, these have been replaced by *super rules*, which include two or more contiguous reactions that form a small pathway of their own. Rules are continually modified to improve the prediction accuracy; *increasing rule stringency* can improve predictions and reduce extraneous choices. The UM-PPS is freely available to all without registration. Its value to the scientific community, for academic, industrial and government use, is good and will only increase.

INTRODUCTION

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 the web, for microbial catabolism of environmental pollutants (1). The intended scope of the UM-BBD was carefully considered. Rather than extracting data randomly from the existing scientific literature, we wished to mitigate bias scientists might have in studying some types of microbial metabolism more than others. For example, since the database contains degradation pathways for benzene, toluene and ethylbenzene, it was less important to cover those of *i*-propylbenzene and *t*-butyl benzene. It was decided that UM-BBD pathways would emphasize the wide range of microbial metabolism of organic functional groups, and its developers would see if this data could be used to predict biodegradation of compounds not found in the database. The first University of Minnesota PredictBT workshop (http://umbbd.msi.umn.edu/predictbt/), to discuss a prototype prediction system, was held in 1998 (2). This led, in 2002, to the development of the web-based University of Minnesota pathway prediction system (UM-PPS, http://umbbd.msi.umn.edu/predict/) (3–4).

The UM-PPS recognizes functional groups in organic compounds that are potential targets of microbial catabolic reactions, and predicts transformations of these groups based on biotransformation rules. Biotransformation rules are based on reactions found in the UM-BBD database or directly in the scientific literature. Periodic PredictBT workshops guide UM-PPS development; a recent one recommended a clearer statement of its scope, which is now found on its home page: UM-PPS predictions are most accurate for compounds that are similar to compounds whose biodegradation pathways are reported in the scientific literature; in environments exposed to air, in moist soil or water, at moderate temperatures and pH, with no competing chemicals or toxins; and which are the sole source of energy, carbon, nitrogen or other essential element for the microbes in these environments, rather than present in trace amounts.

*To whom correspondence should be addressed. Tel: +1 612 635 9122; Fax: +1 612 624 6404; Email: lynda@umn.edu

HARDWARE AND SOFTWARE

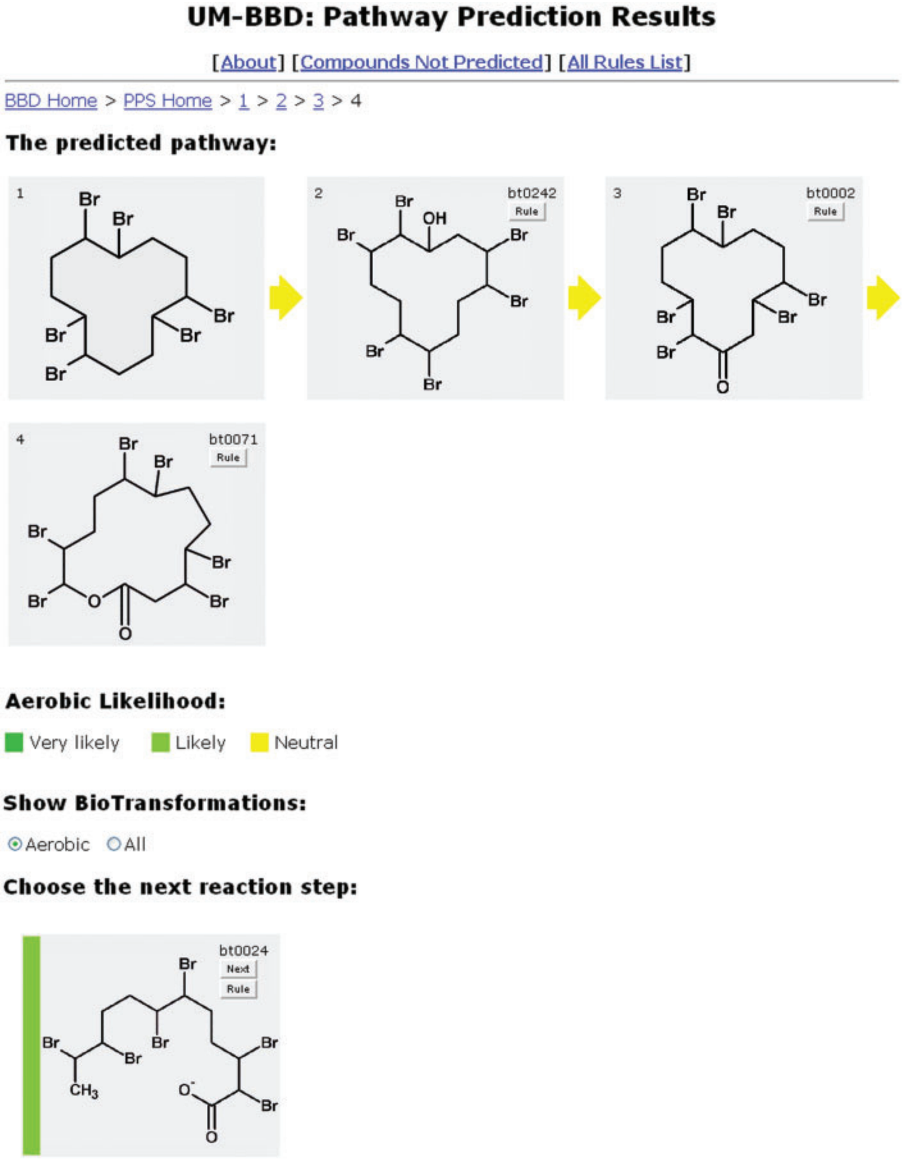
The UM-PPS, and the UM-BBD upon which it is based, are hosted by the Minnesota Supercomputing Institute (http://www.msi.umn.edu/), on an Intel XEON processor using the Linux operating system. The UM-PPS is written primarily in Java and Perl, and uses the JChemBase, Reactor and related ChemAxon modules (http://www.chemaxon.com/) (5), MySQL (http://www.mysql.com/) and the Apache and Tomcat web servers (http://www.apache.org/).

USE AND USAGE

A user enters a query compound into the UM-PPS by drawing it using a MarvinView Java applet (5); the system then creates a SMILES string (6) for it. Alternatively, the user can directly paste or enter a SMILES string. If the compound is one of the UM-PPS's termination compounds, primarily products of common intermediary metabolism, the prediction ends. If not, the string is submitted to all biotransformation rules. If none are triggered, the prediction ends. If one or more rules are triggered, one or more predicted transformations is/are displayed for each rule. The user can choose any of the predicted products to continue the prediction cycle. The predicted pathway after three steps in a representative prediction, for an EU priority pollutant not present in the UM-BBD (hexabromocyclododecane, HBCD), is shown in Figure 1.

COMBINATORIAL EXPLOSION

As more rules were added to the UM-PPS rulebase, users complained of 'too many choices' at each step (1).


UM-BBD: Pathway Prediction Results

[About] [Compounds Not Predicted] [All Rules List]
BBD Home > PPS Home > 1 > 2 > 3 > 4

The predicted pathway:

Aerobic Likelihood:

🟩 Very likely 🟩 Likely 🟨 Neutral

Show BioTransformations:

● Aerobic ○ All

Choose the next reaction step:

Figure 1. A three-step predicted pathway and the choice for the fourth step (ring cleavage) for HBCD, a commercial flame retardant and an EU priority pollutant not found in the UM-BBD. This is a plausible, but perhaps not the most plausible, prediction (see text).

Multiple choices at each step quickly cause combinatorial explosion. This is being addressed in four ways, using aerobic likelihood, relative reasoning, super rules and increased rule stringency.

Aerobic likelihood

Some of the biotransformation rules require molecular oxygen; those represent reactions primarily occurring in bacteria under aerobic conditions, though, in some anaerobic microbes, such enzymes can work under very low partial pressures of oxygen. Others represent strictly anaerobic transformations. Prior to a PredictBT workshop, we asked panelists to rank btrules according to their aerobic likelihood on a five-point Likert scale: Very Likely, Likely, Neutral, Unlikely and Very Unlikely. Two or more independent panelists ranked each rule. The workshop handled those cases that had a wide range of rankings; consensus was reached in all cases. Two panelists agreed to rank future rules. Results are now ordered by color-coded aerobic likelihood. Users can choose to see all predictions or only those more likely under aerobic conditions (1).

Relative reasoning

While aerobic likelihood was an important first step, additional approaches were needed. We introduced relative reasoning, which has been used to improve the prediction of mammalian metabolism (7), to allow certain rules to have priority over others. For example, dioxygenation of an aromatic ring to produce a *vic*-dihydroxy aromatic and extradiol ring cleavage of a *vic*-dihydroxy aromatic ring are both ranked as aerobically likely rules. However, in all 25 UM-BBD compounds where both could theoretically occur, only ring cleavage occurred. This is understandable in biochemical terms, since the goal of oxygenation of an aromatic ring is ring cleavage. Generalizing, all ring cleavage rules were given priority over mono- and dioxygenation of aromatic rings. In other words, if a ring cleavage rule is triggered, triggering of ring oxygenation rules is suppressed. This one relative reasoning rule decreased choices in predictions of aromatic ring degradation by up to 75% (from 16 to 4 for benzyl alcohol) with no loss of sensitivity. Similarly, testing of all rules against the known pathways for all UM-BBD compounds was used to determine other relative reasoning rules; biochemical logic was the final determinant of which were implemented. A list of all biotransformation rules with relative reasoning, and the rules each has priority over, is available: http://umbbd.msi.umn.edu/cgi-bin/relative_reasoning.cgi.

Super rules

While aerobic likelihood and relative reasoning did reduce the number of choices, choosing one-by-one biotransformation steps needed to handle even simple predictions, such as β-oxidation or the degradation of benzenoids, was very time-consuming. Also, at each step it was possible to make a 'wrong' choice and thus not follow a known pathway route. It was decided to combine selected contiguous rules that constitute a small known metabolic pathway into so-called super rules. A super rule takes the starting compound and converts it directly to the end product of the small pathway. It may also handle all or some of the pathway intermediates. The first super rule was for β-oxidation. There are several super rules for benzenoid degradation; the one handling the ring cleavage product of the intradiol (ortho) degradation pathway for a single aromatic ring is shown in Figure 2a and b. Descriptions for super rules are indicated by (*) on the list of all rules: http://umbbd.msi.umn.edu/servlets/pageservlet?ptype = allrules.

Increased rule stringency

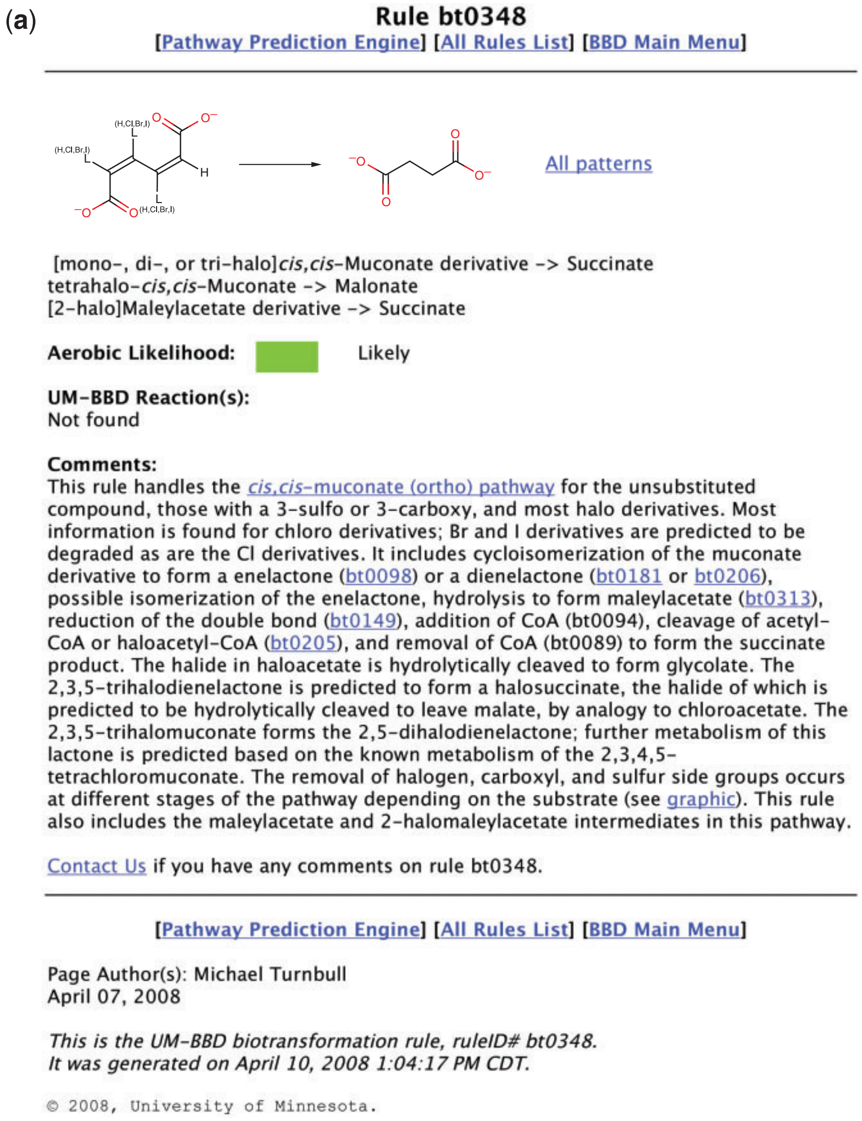
It is difficult to decide on rule stringency. On one hand, the degradation of a specific UM-BBD compound would be correctly predicted by writing a very specific rule for each of its reactions and using relative priority to remove any other triggered rules. The UM-PPS rulebase could be a set of these specific rules for all UM-BBD compounds. Such an overspecified system, however, would not correctly predict the degradation of compounds not found in the UM-BBD. On the other hand, a very general rule, such as the original rules for hydrolytic dehalogenation (bt0022) or decarboxylation (bt0051), which applied indiscriminately to any kind of C−X (X = F, Cl, Br, I) or C–COOH bond, respectively, will frequently be inappropriately triggered. We improved these rules, and others, by increasing their stringency.

Hydrolytic dehalogenation (bt0022), aerobically neutral and originally very general, was restricted to halogens on methyl groups in linear aliphatic compounds. The C–X bond in halogenated aromatics is more stable and often only broken after ring cleavage (as shown in Figure 2); this was implemented in the UM-PPS, though there are exceptions. To handle one of them, a specific rule (bt0366) was written for 4-halophenylacetate 3,4-dioxygenase (EC 1.14.12.9). Similarly, halogens in cyclic aliphatics often remain on the ring until ring cleavage; this was also implemented in the UM-PPS, though, again, there are exceptions. Reductive dehalogenation (bt0029), aerobically unlikely, remains a general rule.

Decarboxylation (bt0051), aerobically neutral and originally a very general rule, was rewritten to only decarboxylate structures that undergo decarboxylation in the UM-BBD, or are known to be decarboxylated by microbial enzymes as reported in enzyme databases (EC 4.1.1.-). Such structures include most 2-oxo-, 2-amino-, 2-hydroxy-, 2-aryl-, 3-oxo-, 3-lactone, 3-ene-5-oxo, 3-hydroxy, and 3-aryl- carboxylates and some 2-*N*-heterocyclic aromatic carboxylates. The rule was restricted from acting on substrates for β-oxidation (bt0337).

VALIDATION

Validation of such a system is not trivial. At a minimum, it should give plausible predictions for the UM-BBD compounds on which it is based. Second, it should plausibly handle compounds not found in the UM-BBD. In both cases, measuring plausibility is problematic; it is not synonymous with known metabolism, because the universe of environmental microbial catabolism is not well-sampled.



Rule bt0348

[Pathway Prediction Engine] [All Rules List] [BBD Main Menu]

All patterns

[mono-, di-, or tri-halo]*cis,cis*-Muconate derivative -> Succinate
tetrahalo-*cis,cis*-Muconate -> Malonate
[2-halo]Maleylacetate derivative -> Succinate

Aerobic Likelihood: 🟩 Likely

UM-BBD Reaction(s):
Not found

Comments:
This rule handles the *cis,cis*-muconate (ortho) pathway for the unsubstituted compound, those with a 3-sulfo or 3-carboxy, and most halo derivatives. Most information is found for chloro derivatives; Br and I derivatives are predicted to be degraded as are the Cl derivatives. It includes cycloisomerization of the muconate derivative to form a enelactone (bt0098) or a dienelactone (bt0181 or bt0206), possible isomerization of the enelactone, hydrolysis to form maleylacetate (bt0313), reduction of the double bond (bt0149), addition of CoA (bt0094), cleavage of acetyl-CoA or haloacetyl-CoA (bt0205), and removal of CoA (bt0089) to form the succinate product. The halide in haloacetate is hydrolytically cleaved to form glycolate. The 2,3,5-trihalodienelactone is predicted to form a halosuccinate, the halide of which is predicted to be hydrolytically cleaved to leave malate, by analogy to chloroacetate. The 2,3,5-trihalomuconate forms the 2,5-dihalodienelactone; further metabolism of this lactone is predicted based on the known metabolism of the 2,3,4,5-tetrachloromuconate. The removal of halogen, carboxyl, and sulfur side groups occurs at different stages of the pathway depending on the substrate (see graphic). This rule also includes the maleylacetate and 2-halomaleylacetate intermediates in this pathway.

Contact Us if you have any comments on rule bt0348.

[Pathway Prediction Engine] [All Rules List] [BBD Main Menu]

Page Author(s): Michael Turnbull
April 07, 2008

This is the UM-BBD biotransformation rule, ruleID# bt0348.
It was generated on April 10, 2008 1:04:17 PM CDT.

© 2008, University of Minnesota.

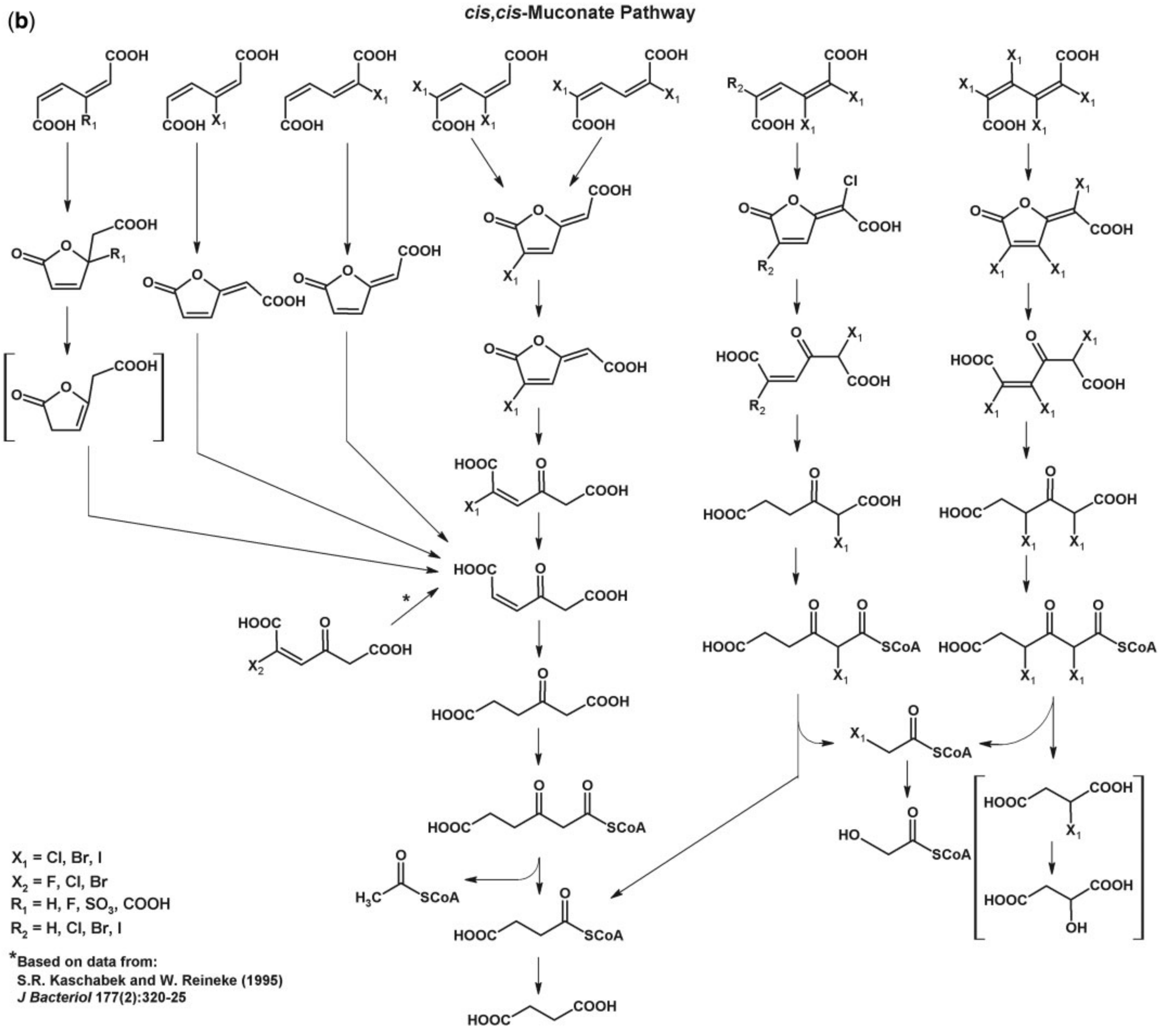
Figure 2. (**a**) The benzenoid *cis,cis*-muconate, intradiol (ortho) lower cleavage pathway super rule (bt0348). The complete rule is found at: http://umbbd.msi.umn.edu/servlets/rule.jsp?rule = bt0348. (**b**) The graphic for the benzenoid *cis,cis*-muconate, intradiol (ortho) lower cleavage pathway super rule (bt0348) shown in (a).

Initially, when UM-BBD reactions were assigned to individual rules, the UM-PPS was validated by how many of the known UM-BBD pathways could be predicted, starting from the initial compound. Over 98% (111/113) of such pathways could be predicted (4). However, there were many, sometimes very many, extraneous predictions.

After implementation of aerobic likelihood and relative reasoning, but before major use of super rules or increasing rule stringency, UM-PPS predictions for the first transformation step were validated using two compound sets with known first product(s) of biodegradation: a randomly selected set of 50 UM-BBD compounds, and 25 pesticides not in the UM-BBD. Sensitivity [TP/(TP + FN)] and selectivity (TP/N) were calculated, where TP = number of known compounds predicted, FN = number of known compounds not predicted and N = total number of predictions. Sensitivity was 0.74 for both sets; selectivity was 0.16 for the UM-BBD set and 0.17 for the pesticide set. The names, CAS numbers and structures of the 25 compounds in the pesticide set are available in Supplementary Material.

OTHER SYSTEMS

There are no other free microbial catabolic pathway prediction tools available to the public. CATABOL and META-CASE are commercial systems (8,9). A system that uses similarity searching to improve biodegradation pathway prediction has been reported, but is not available for public use (10). With the help of Lhasa Limited, we are developing a stand-alone, proprietary system based on the UM-PPS, provisionally called MEPPS (1). The free UM-PPS and proprietary MEPPS will coexist in synchrony for the foreseeable future.



X_1 = Cl, Br, I
X_2 = F, Cl, Br
R_1 = H, F, SO_3, COOH
R_2 = H, Cl, Br, I

*Based on data from:
S.R. Kaschabek and W. Reineke (1995)
J Bacteriol 177(2):320-25

Figure 2. Continued.

DISCUSSION AND CONCLUSIONS

The UM-PPS was validated for single steps in a biodegradation pathway prediction. While single-step validation sensitivity is satisfactory, selectivity is rather low: the UM-PPS often predicts transformation products in addition to those that have been observed experimentally or under real world conditions. However, in pure culture studies, on which most scientific literature is based, microbes will tend to express the most thermodynamically efficient pathways. Under environmental conditions, the product spectrum is broader, since it depends more on the available enzyme pool. Thus, a system with this level of sensitivity and selectivity is of value.

Each rule is presently assigned one aerobic likelihood value. However, there is need for variable aerobic likelihood based on the chemical structure of the substrate. For example, the aerobic degradation pathway of the highly halogenated compound shown in Figure 1 is plausible, but it is probably aerobically unlikely, not neutral; this is supported by its known anaerobic degradation (11). If the aerobic likelihood for rules were changed to ‘unlikely’ for such a compound, anaerobic dehalogenation would be shown as an alternative choice, since it is also aerobically unlikely. This would result in more plausible prediction steps, where multiple dehalogenation and Baeyer–Villiger oxidation reactions would occur in unknown sequence.

Structure-dependent aerobic likelihoods are also needed for some super rules, such as the one shown in Figure 2. The large range of starting compounds handled by this rule means it could have likely, neutral or unlikely aerobic likelihood. For example, it is known that in a series of compounds with the same carbon skeleton, the likeliness of aerobic biodegradation is inversely proportional to the number of halogens it contains.

Super rules also have limitations. The steroid cholesterol and testosterone biodegrade by similar pathways. A super rule was written that takes the common steroid intermediate for both compounds through its first ring cleavage. However, testosterone needs one initial step to be transformed into this intermediate, cholesterol needs several others, and though it could be predicted that other steroids would use that same intermediate, each would need a variable number and type of biotransformations to reach that point. It would be desirable for the UM-PPS to 'look ahead' to such readily degradable intermediates, and guide the user to biotransformations that would lead to it.

The UM-PPS is fully functional and freely available to all without registration. Ranking rules by aerobic likelihood, use of relative reasoning, implementing super rules and increasing stringency of certain rules have improved its predictions, and more improvements are planned. Over 150 unique compounds are now entered each month. It presently provides good value to the scientific community, for academic, industrial and government use, and its value will only increase.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Rachael Long and Michael Turnbull for creating many rules and super rules for the UM-PPS; and the participants of the PredictBT workshops for their guidance. This work was supported by U.S. National Science Foundation (NSF9630427 to L.W. and L.E.); Swiss National Science Foundation (PA002-113140 to K.F.); Lhasa Limited (to L.E. and L.W.); the Minnesota Supercomputing Institute. Funding to pay the Open Access publication charges for this article was provided by US National Science Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Ellis,L.B.M., Roe,D. and Wackett,L.P. (2006) The University of Minnesota biocatalysis/biodegradation database: the first decade. *Nucleic Acids Res.*, **34**, D517–D521.
2. Wackett,L.P., Ellis,L.B.M., Speedie,S., Hershberger,C.D., Knackmuss,H.-J., Spormann,A., Walsh,C.T., Forney,L.J., Punch,W.F., Kazic,T. *et al.* (1999) The prediction of microbial biodegradation and biocatalysis. *Amer. Soc. Microbiol. News*, **65**, 87–93.
3. Hou,B.K., Wackett,L.P. and Ellis,L.B.M. (2003) Predicting microbial catabolism: a functional group approach. *J. Chem. Inf. Comp. Sci.*, **43**, 1051–1057.
4. Hou,B.K., Ellis,L.B.M. and Wackett,L.P. (2004) Encoding microbial metabolic logic: predicting biodegradation. *J. Ind. Microbiol. Biotechnol.*, **31**, 261–272.
5. Csizmadia,F. (2000) JChem: Java applets and modules supporting chemical database handling from web browsers. *J. Chem. Inf. Comp. Sci.*, **40**, 323–324.
6. Weininger,D. (1988) SMILES: a chemical language for information systems. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31–36.
7. Button,W.G., Judson,P.N., Long,A. and Vessey,J.D. (2003) Using absolute and relative reasoning in the prediction of the potential metabolism of xenobiotics. *J. Chem. Info. Comp. Sci.*, **43**, 1371–1377.
8. Dimitrov,S., Kamenska,V., Walker,J.D., Windle,W., Purdy,R., Lewis,M. and Mekenyan,O. (2004) Predicting the biodegradation products of perfluorinated chemicals using CATABOL. *SAR QSAR Environ. Res.*, **15**, 69–82.
9. Klopman,G. and Tu,M. (1997) Structure-biodegradability study and computer-automated prediction of aerobic biodegradation of chemicals. *Environ. Toxicol. Chem.*, **16**, 1829–1835.
10. Oh,M., Yamada,T., Hattori,M., Goto,S. and Kanehisa,M. (2007) Systematic analysis of enzyme-catalyzed reaction patterns and prediction of biodegradation pathways. *J. Chem. Inf. Modeling*, **47**, 1702–1712.
11. Gerecke,A.C., Giger,W., Hartmann,P.C., Heeb,N.V., Kohler,H.P., Schmid,P., Zennegg,M. and Kohler,M. (2006) Anaerobic degradation of brominated flame retardants in sewage sludge. *Chemosphere*, **64**, 311–317.