D488–D491 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 18 September 2009*
doi:10.1093/nar/gkp771

The University of Minnesota Biocatalysis/ Biodegradation Database: improving public access

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

[1]Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN 55455 and [2]Department of Biochemistry, Molecular Biology and Biophysics, University of Minnesota, St Paul, MN 55108, USA

Received August 20, 2009; Revised August 31, 2009; Accepted September 1, 2009

ABSTRACT

The University of Minnesota Biocatalysis/ Biodegradation Database (UM-BBD, http://umbbd .msi.umn.edu/) began in 1995 and now contains information on almost 1200 compounds, over 800 enzymes, almost 1300 reactions and almost 500 microorganism entries. Besides these data, it includes a Biochemical Periodic Table (UM-BPT) and a rule-based Pathway Prediction System (UM-PPS) (http://umbbd.msi.umn.edu/predict/) that predicts plausible pathways for microbial degradation of organic compounds. Currently, the UM-PPS contains 260 biotransformation rules derived from reactions found in the UM-BBD and scientific literature. Public access to UM-BBD data is increasing. UM-BBD compound data are now contributed to PubChem and ChemSpider, the public chemical databases. A new mirror website of the UM-BBD, UM-BPT and UM-PPS is being developed at ETH Zürich to improve speed and reliability of online access from anywhere in the world.

INTRODUCTION

The University of Minnesota Biocatalysis/Biodegradation Database (UM-BBD, http://umbbd.msi.umn.edu/) is a free online database that catalogues information on microbial biocatalytic reactions and biodegradation pathways, primarily for xenobiotic organic compounds. It has served as an important online reference tool for biocatalysis and biodegradation pathways for over 14 years. Currently, it contains almost 1200 compounds, over 800 enzymes, almost 1300 reactions, almost 500 microorganism entries and almost 200 pathways. It has about 200 000 unique users per year. We have previously reported its content, methods and the major changes during its first decade (1–5).

Besides pathway, compound, enzyme and reaction data, the UM-BBD also includes a Pathway Prediction System (UM-PPS; described below) and a Biochemical Periodic Table (UM-BPT, http://umbbd.msi.umn.edu/periodic/). UM-BBD information is based on the scientific literature that describes microbial metabolic pathways in molecular detail. However, much more information on microbial metabolism is available in the scientific literature in less detail, including descriptions of microbial transformation of chemical elements. As described previously (5), the UM-BPT was developed to indicate the wide range of these microbial transformations. It includes biological information for almost all stable, non-noble-gas elements.

Based on the information found in the UM-BBD, the UM-PPS (http://umbbd.msi.umn.edu/predict/) predicts microbial catabolism of organic compounds. Since the system was created in 2002, its biotransformation rule base has grown to 260 entries. A new version of the graphical user interface (GUI) offers an improved visual design and functionality. The system infrastructure was improved to allow additional metabolic logic to improve pathway prediction results. Its predicting speed was nearly doubled. These changes lead to a smarter and faster UM-PPS that is used up to 1000 times each month.

DATABASE UPDATES AND CHANGES

The knowledge base of UM-BBD grew overall about 30% since its last report in September, 2005 (5). Over 300 new reactions, over 300 new compounds and over 200 new enzymes contributed to the increase. UM-BBD pathways have grown by 50, to almost 200, over the last 4 years, and there have been 25 updates of existing pathways.

In 2007, the UM-BBD was moved to a new UNIX server with high-end hardware architecture located in the Minnesota Supercomputing Institute (http://www .msi.umn.edu/). The database's URL was changed to http://umbbd.msi.umn.edu/ at that time.

The UM-BBD is connected to many other web resources; these links must be continually maintained. In the past four years, on compound, enzyme and reaction pages, links have been updated to BRENDA (6) and

*To whom correspondence should be addressed. Tel: +1 612 625 9122; Fax: +1 612 624 6404; Email: lynda@umn.edu

PubMed (7), added to PubChem (8) and IntEnz (9) and changed for enzyme structures to the NCBI Structure database (7) and for microorganisms to the World Data Centre for Microorganisms (http://wdcm.nig.ac.jp/).

IMPROVED PUBLIC ACCESS

Since 2006, UM-BBD compounds have been deposited in the PubChem compound database (8). Molecular models of all UM-BBD compounds are now available from PubChem. A user may access these models by following the 'PubChem Substance Entry' link on each compound page. Compound information is downloadable at the bottom of each PubChem Substance (and Compound) entry. Since 2008, UM-BBD compounds have also been deposited in the ChemSpider compound database (10).

The UM-BBD was first mirrored at the European Bioinformatics Institute in 2000. This mirror site has been in synchrony with the UM-BBD for the last 9 years, but is limited to information on UM-BBD compounds, reactions, enzymes and pathways. A second mirror site for the UM-BBD is being developed at ETH Zürich under the direction of Dr. Kathrin Fenner. The new site will completely mirror the UM-BBD, including the UM-BPT and UM-PPS. The need for such a complete mirror is indicated by web use statistics: the four most popular pages, in rank order, are home, search, UM-PPS home and links.

IMPROVED UM-PPS

The UM-PPS (11) is becoming a widely recognized pathway prediction system used to improve bioremediation (12). Predictions are based on biotransformation rules that, in turn, are derived from reactions found in the UM-BBD and the scientific literature. The UM-PPS most accurately predicts compounds that are similar to compounds with known biodegradation mechanisms, for microbes under aerobic conditions and when the compounds are the sole source of energy, carbon, nitrogen or other essential elements for these microbes. For more information, see 'About the PPS web page' (http://umbbd.msi.umn.edu/predict/aboutPPS.html).

The improved GUI

The UM-PPS provides a user-friendly GUI that allows a user to browse the prediction results as shown in Figure 1. The UM-PPS predicts one or more biotransformations if one or more biotransformation rules can be triggered by a user's query compound. A biodegradation pathway consisting of a series of compounds will be shown based on a user's choices during all intermediate steps. In a predicted pathway, compounds are connected by coloured arrows. Each coloured arrow represents a biotransformation with a specific aerobic likelihood value, from very likely to very unlikely. Aerobic likelihood evaluation and assignment has been described (11) and a legend is available under the pathway. Below that legend, one or more predicted compounds may be displayed if the last compound in the predicted pathway still triggers one or more rules.

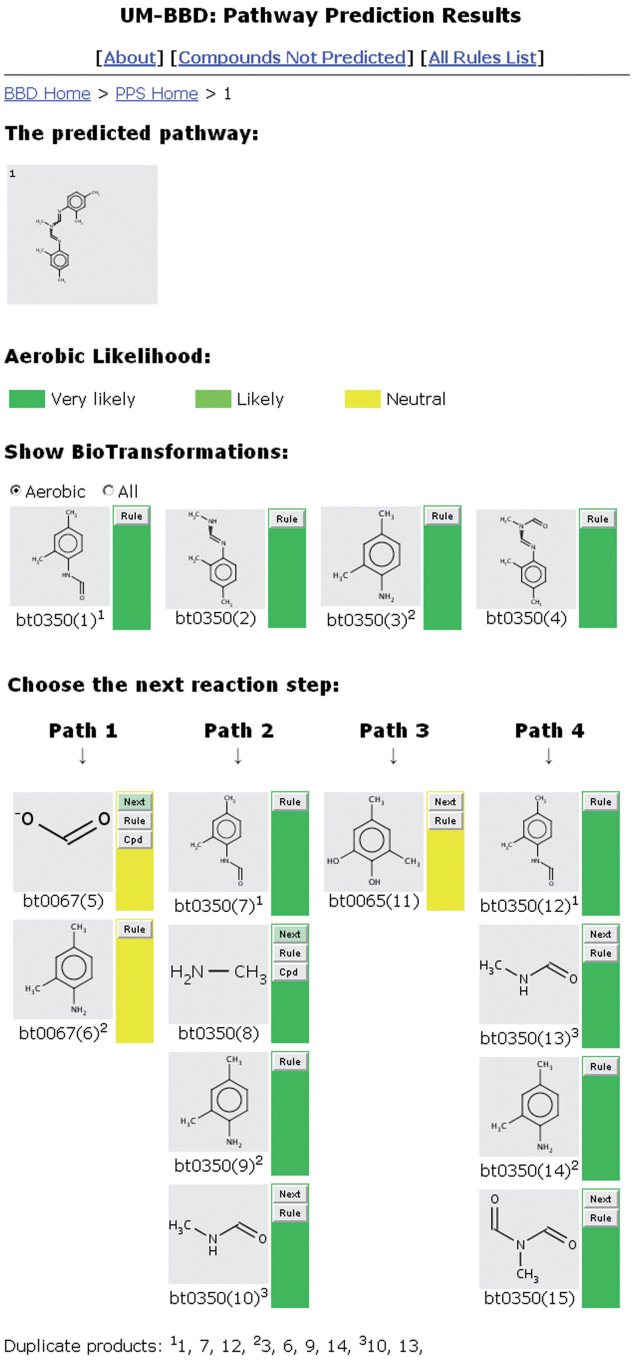

Figure 1. The first two-step prediction for the pesticide amitraz (CC1=CC(=C(C=C1)N=CN(C)C=NC2=C(C=C(C=C2)C)C)C).

A new green 'next' button may be presented at the top-right corner of a compound box if this compound is an end node of that pathway. A navigation bar is provided above a predicted pathway where a user may review system's suggested compounds and revise his or her choices in any previous steps.

The improved GUI now allows users to view a two-step prediction as an optional choice. The web page layout is similar to the one-step prediction previously described (5). Duplicate compounds are only shown in two-step

predictions; one-step predictions remove all such compounds that will cause unwanted prediction loops. Duplicate compounds are marked by superscripts and grouped under a series of incremental numbers in a bottom note.

Multistep prediction may enable better analysis of best choices. For example, in the two-step prediction of the pesticide amitraz (Figure 1), there are four very likely predicted compounds produced by two cleavage biotransformations in the first step. In the second step, Compound 1 (2,4-dimethylaniline) goes through a neutral biotransformation to Compound 5 that would be predicted to accumulate in the environment. If a user traces the largest piece in the predicted pathway, Compound 3 is predicted to degrade to Compound 11 (2,4-dimethylaniline) in Path 3. Since Compounds 1 and 11 are duplicate compounds, Path 3 merges with Path 1. Similarly, Paths 2 and 4 merge to Path 1, either directly via duplicate compound Group 1 or indirectly following Path 3 via duplicate compound Group 2. The above analysis shows that all four predicted pathway connect substrate amitraz to its known product 2,4-dimethylaniline. Path 1 is the shortest pathway. Based on the above, Path 1 may be the best candidate pathway to continue the prediction.

Implementation of metabolic logic

Immediate feature. As the number of rules increases, the UM-PPS predicts more products at each step. In a multiple-step pathway prediction, this can easily cause combinatorial explosion, a problem that has been addressed previously (13). To solve this problem, additional metabolic logic has been applied to remove implausible products in the predicted pathways. The system first introduced an immediate feature that guides users to most likely pathways as soon as possible by skipping one or more transient steps without manual intervention. This feature is only activated when the 'aerobic' option is chosen, and it is only applicable to all very likely and some likely biotransformations. If a user chooses to see all biotransformations regardless of the aerobic likelihood values, the immediate feature is turned off.

Relative reasoning. The UM-PPS uses relative reasoning as another effective approach to limit combinatorial explosion. To allow this feature, a relative reasoning field was added to the rule table, which can give certain rules priority over others. One relative reasoning entry decreased choices in prediction of aromatic ring degradation to 75% with no loss of sensitivity (11). As the number of relative reasoning entries grows, the priority relationship among rules becomes complex. A validation function was developed to examine and report any potential deadlock in existing entries (e.g. Rule 1 has priority over Rule 2, Rule 2 over Rule 3 and Rule 3 over Rule 1).

Super rules. Besides the relative reasoning, the improved UM-PPS infrastructure allows super rules to provide additional reduction to the combinatorial explosion. A super rule was described as a combination of selected contiguous rules that form a small known metabolic pathway of its own. These rules can handle more than one biotransformation intermediates, guide users to some known pathways, such as β-oxidation, quickly and significantly shorten the prediction processes.

Variable aerobic likelihood. Since the UM-PPS uses a limited number of rules to make predictions of a wide range of different compounds, the aerobic likelihood of some rules is not always as accurate as expected. To make more plausible predictions, the system introduced variable aerobic likelihood values. To implement this feature, we assigned regular expression patterns to selected rules. During a prediction process, structure-based aerobic likelihood values will be shown if a substrate triggers one or more rules and matches their patterns. This feature gives more accurate likelihood for rules triggered by substrates with certain chemical structures.

Summary. In August, 2009, there were 259 total rules, 19 of them with immediate feature, 115 of them with relative reasoning entries, 13 of them being super rules and 27 with variable aerobic likelihood entries.

Improved compatibility and speed

As UM-PPS becomes a widely used tool, its compatibility with a large number of browsers and prediction speed become more and more important. The former UM-PPS used Java applets (14) to display predicted compounds. Since the applets cannot start until the Java Virtual Machine is running, users often had significant start-up time before display of the prediction, or might even not see any compound graphics if their web browsers did not support a Java plug-in. Now, the UM-PPS uses static Portable Network Graphics (PNG), which saves graphics loading time and allows the system to be compatible with more web browsers. At the same time, a user can also view compounds within a Java applet in a pop-up window, and take advantage of its many functions, by simply clicking on the compound PNG.

A user submits a query compound to all biotransformation rules and the UM-PPS makes prediction; this procedure is called a prediction cycle. The former UM-PPS matched the query compound with all rules one by one, which is a time-consuming process on a multistep prediction. Now, the UM-PPS is hosted at a high-end 4 CPU server, and the upgraded hardware permits improvement to the system's computing performance, using concurrent computing strategy. During a prediction cycle, all rules are divided evenly to four groups and each group runs against the query compound in its own task. These four subtasks are running concurrently, and predicted results are merged together when the last subtask is done. The performance is only twice, not four times, as fast, due to additional system overhead and use of a server shared with other web applications.

CONCLUSIONS

The UM-BBD has continued its growth in the past 4 years, and its influence increases with over 200 000 unique users per year. Access to its data is improved through contributions of its compounds to public chemical databases. The UM-PPS is also actively growing, and a new mirror at ETH Zürich that will include the UM-BBD, UM-BPT and UM-PPS will soon improve access to the world.

ACKNOWLEDGEMENTS

The authors thank the May 2007 and 2009 PredictBT workshop participants for their guidance on improving the UM-BBD and UM-PPS. They thank the students in BioC/MicE 5309 at the University of Minnesota for starting many UM-BBD pathways. They thank Michael Turnbull for creating many rules for the UM-PPS and other web work. They thank Dr Kathrin Fenner and Mr Peter Bircher at ETH Zürich for providing the space and considerable effort required for the mirror site.

FUNDING

US National Science Foundation (NSF0543416) and the Minnesota Supercomputing Institute. Funding for open access charge: US National Science Foundation (NSF05434161).

Conflict of interest statement. None declared.

REFERENCES

1. Ellis,L.B., Hershberger,C.D. and Wackett,L.P. (1999) The University of Minnesota Biocatalysis/Biodegradation database: specialized metabolism for functional genomics. *Nucleic Acids Res*, **27**, 373–76.
2. Ellis,L.B., Hershberger,C.D. and Wackett,L.P. (2000) The University of Minnesota Biocatalysis/Biodegradation database: microorganisms, genomics and prediction. *Nucleic Acids Res*., **28**, 377–379.
3. Ellis,L.B., Hershberger,C.D., Bryan,E.M. and Wackett,L.P. (2001) The University of Minnesota Biocatalysis/Biodegradation Database: emphasizing enzymes. *Nucleic Acids Res*., **29**, 340–343.
4. Ellis,L.B., Hou,B.K., Kang,W. and Wackett,L.P. (2003) The University of Minnesota Biocatalysis/Biodegradation Database: post-genomic data mining. *Nucleic Acids Res*., **31**, 262–265.
5. 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.
6. Chang,A., Scheer,M., Grote,A., Schomburg,I. and Schomburg,D. (2009) BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. *Nucleic Acids Res*., **37**, D588–D592.
7. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al*. (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **37**, D5–D15.
8. Wang,Y., Xiao,J., Suzek,T.O., Zhang,J., Wang,J. and Bryant,S.H. (2009) PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res*., **37**, W623–W633.
9. Fleischmann,A., Darsow,M., Degtyarenko,K., Fleischmann,W., Boyce,S., Axelsen,K.B., Bairoch,A., Schomburg,D., Tipton,K.F. and Apweiler,R. (2004) IntEnz, the integrated relational enzyme database. *Nucleic Acids Res*., **32**, D434–D437.
10. Williams,A.J. (2008) Internet-based tools for communication and collaboration in chemistry. *Drug Discov. Today*, **13**, 502–506.
11. Ellis,L.B.M., Gao,J., Fenner,K. and Wackett,L.P. (2008) The University of Minnesota pathway prediction system: predicting metabolic logic. *Nucleic Acids Res*., **36**, W427–W432.
12. De Lorenzo,V. (2008) Systems biology approaches to bioremediation. *Curr. Opin. Biotechnol*., **19**, 579–589.
13. Fenner,K., Gao,J., Kramer,S., Ellis,L.B.M. and Wackett,L.P. (2008) Data-driven extraction of relative reasoning rules to limit combinatorial explosion in biodegradation pathway prediction. *Bioinformatics*, **24**, 2079–2085.
14. Csizmadia,F. (2000) J Chem: Java applets and modules supporting chemical database handling from web browsers. *J. Chem. Inf. Comput. Sci*., **40**, 323–324.