Biodegradation Prediction Tools

Florencio Pazos and Víctor de Lorenzo

Abstract

Experimental approaches for determining the environmental fate of new molecules generated by the modern chemical and pharmaceutical industry cannot cope with the pace at which new of these substances are synthesized, thus raising questions on their ultimate fate if released into the environment. This has fostered the development of different web-based and publicly available platforms that deliver an appraisal of the amenability of given chemical species to microbial biodegradation. One major class of such predictors foretell the final destiny of an input chemical formula, either as an end-point state (i.e., degradable or not) or as an estimation of half-life under certain circumstances. These tools are characteristically automated and thus most suitable for screening large collections of compounds without any expert knowledge. A second type of platforms provides information on the possible – often alternative – biodegradation routes that the compounds under examination may go through, with an indication of possible intermediates and enzymatic reactions. Such pathway-based tools require a degree of interactivity by the user and are more suited to analyses of individual target molecules. The protocols detailed below describe the practical usage of one platform of each type, specifically, EAWAG-PPS (formerly UM-PPS) and *BiodegPred*. Both take as an input the formulas of the compounds, but they deliver different and somewhat complementary information on their most likely environmental fate.

Keywords: Biodegradation, Bioremediation, Metabolism, Prediction, Web server

1 Introduction

The large collection of molecules produced in bulk amounts by the chemical and pharmaceutical industries since the onset of industrialization in the nineteenth century is at the basis of our Western societies and economies. Alas, the same industrial activities typically release into the environment vast quantities of chemical products that can be harmful for the ecosystem. Toxic waste is often generated as a side product of the synthesis or utilization of a molecule of interest, and it becomes noxious once it is released to the environment – either accidentally or deliberately. While large loads of harmful contaminants (e.g., oil, heavy metals) can be partially coped with through mere physicochemical methods (e.g., intensive in-source treatment, physical removal, landfilling), the most typical cases of

T.J. McGenity et al. (eds.), *Hydrocarbon and Lipid Microbiology Protocols*, Springer Protocols Handbooks (2016) 115–124,
DOI 10.1007/8623_2015_87, © Springer-Verlag Berlin Heidelberg 2015, Published online: 17 May 2015

pollution are those in which the levels of the toxic molecules are low enough to make mechanical removal inefficient while being high enough to cause a distinct environmental impact, often on the long run. Some of the compounds at stake are naturally degraded by environmental physicochemical abiotic processes (photolysis, oxidation, etc.) which transform molecules with a given toxicity into less harmful products. Other molecules can be totally or partially metabolized (or co-metabolized) by environmental microorganisms. But many other substances, termed "recalcitrant," are not "removed" by any of these means, and they remain in the afflicted sites for very long periods of time.

Biodegradation is the ability acquired by certain environmental organisms to catabolize compounds that do not form part of the standard central metabolism. The driving forces for the emergence of such abilities include both the advantage of benefiting from unusual carbon sources and the counteracting of their chemical toxicity [1, 2]. The rational exploitation of biodegradative capacities of naturally occurring or recombinant organisms (generally microorganisms) for removing chemicals from the environment (in particular in cases of low-level but extensive pollution) is generically called "bioremediation" [3, 4]. This approach has advantages and drawbacks. In one hand, since biodegradation routes are evolved or designed to use the target compound as a carbon/energy source, it generally gets completely "mineralized" (i.e., transformed into CO_2, H_2O, and inorganic small ions), which is more desirable than a partial transformation such as that generally associated to abiotic degradation. On the other hand, releasing bacteria (eventually with genetic modifications) that may be able to survive in the environment competing with others raises a large number of issues [5].

Determining the environmental fate of a new chemical before releasing it to the external medium is crucial for designing appropriate strategies for its synthesis, handling, and disposal or even avoiding its usage/release at all. Strict normatives at national and supranational levels control the procedures for determining the environmental fate of substances and the criteria for allowing their usage or not depending on the results of these procedures (e.g., Williams et al. [6]). It is easy to grasp that gathering experimentally enough data on the fate of each of many molecules that are produced every day by synthetic chemists is very consuming in terms of time and resources. Typical tests involve releasing the compound in a controlled environment and measure its concentration in forthcoming samples taken over long periods of time to determine the kinetics of its eventual degradation (e.g., the "half-life" time required to reduce the concentration to a half of the original one). An additional problem is that measuring the disappearance of the original product is not enough, since intermediates of the degradative pathway (not targeted by the measurement) can be hazardous too. Meanwhile, new chemicals are designed at a pace that cannot be

coped by these *wet* procedures. For these reasons, the in silico prediction of the biodegradative feasibility of a given chemical compound in the environment is of crucial importance since it could help restricting the experimental time/resources devoted to the task [7, 8]. Indeed, the "benign by design" concept, i.e., take into account the (predicted) proneness to degradation as a positive aspect when designing a molecule, is getting more popular in the chemical field [9].

From a methodological point of view, predicting the biodegradative potential of a compound from its chemical structure is, in essence, similar to predicting any other property, such as its melting point, water solubility, or organismal toxicity. The prediction of toxicity is a much more studied issue due to the more direct relationship with human health and the higher difficulty in performing the experiments: i.e., predicting the toxicity of new drugs in humans. On the contrary, predicting biodegradation is a less-explored subject. In principle this is a more difficult task since it depends on many factors apart from the chemical structure of the compound, such as the physicochemical and biological characteristics of a particular environment: water/soil, pH, microbial communities present, etc. Most attempts for predicting toxicity are based, in one way or another, on "quantitative structure-activity relationships" (QSAR) approaches. Some biodegradability predictors also use these general concepts, while also ad hoc strategies were specifically designed to this particular problem.

Taking into account the output they produce, existing biodegradation predictors can be classified in two main classes. In one hand, there are platforms which only predict the final fate of a given compound, either in a quantitative way (to which extent the molecule under examination is going to be degraded, "half-life", etc.) or in a qualitative manner: whether the chemical species at issue is going to be degraded or not (according to some criteria). The second type of predictors includes methods which, apart from predicting the final fate, provide some information on the biodegradative pathway it goes through and the intermediate/final products of the process. Both approaches have pros and cons. While the second class of methods provides more information on the degradation process, they also require in many cases some interactivity or additional input from the user. On the contrary, the methods within the first group are more automatic and hence amenable for application to large collections of compounds without user intervention or expert knowledge.

The freely available alternatives within the first group include, for example, the BIOWIN system which, together with other predictors of different properties of molecular structures, is incorporated in the *EPI Suite* distributed by the US Environmental Protection Agency (EPA) (*see* **Note 1**). BIOWIN is based on regression models where compounds are described by vectors

coding mainly for the occurrence of substructures in the molecule. Several models for predicting biodegradation are contained in BIOWIN, which differ in the criteria used for defining biodegradability (based on different normatives/databases), the scenarios for biodegradation they were designed for (e.g., hydrocarbon degradation, methanogenic anaerobic degradation, etc.), and the output they produce (qualitative or quantitative). Another example of this class of platforms is the BDPServer [10] which uses a machine learning system ("decision trees") fed with a description of the molecules as vectors coding for the frequency of atom triplets plus molecular weight and water solubility if available. For training the system, the environmental fate of the compounds present in the UM-BBD [11] (*see* **Notes 2** and **4**) was in silico inferred based on whether a pathway connecting them with the central metabolism can be found with the information available at that database. This system has been updated to include other molecular descriptors, other machine learning systems and, more importantly, training sets based on "real" experimental biodegradation data (see below). Indeed, this new version (BiodegPred) is conceived as a "multipredictor": the user can run his/her compound against three different biodegradability predictors (based on three different criteria/databases) as well as a toxicity predictor.

Within the second group, the Pathway Prediction System of the UM-BBD (UM-PPS) allows to interactively infer not only the final environmental fate of a compound but the possible route(s) for its degradation and the intermediates involved as well [12]. The system is based on a set of chemical transformations of functional groups frequently observed in biodegradation processes (called "rules"). These rules are applied to the functional groups found in the compound entered by the user, leading to a number of possible virtual products. The process is iterated for these products until the resulting compounds can enter into the central metabolism and/or no additional rules apply for them. The process is also interactive since the user can choose, from the eventual many alternative routes, which ones to go on exploring, defining in this way the complete biodegradation pathway for the initial compound. This system has been recently improved with a machine learning approach that, trained with known examples of biodegradation, allows assigning probabilities to the different pathways [13]. A similar concept is used in the PathPred system [14] of the KEGG metabolic resource [15]. It uses a set of transformations between molecular substructures (called "rpairs") which are less specific than the transformations of functional groups used in UM-PPS, since they involve smaller molecular fragments. Consequently, the main difference is that PathPred generates many more possible compound conversions. Indeed, this approach is generic for "predicting" metabolic transformations and its application to biodegradation involves mainly using it with the "rpair" transformations frequently observed

in KEGG's "xenobiotic biodegradation" pathways. Another approach is followed by the CATABOL/CATALOGIC software [16, 17]. In these systems, the biodegradation pathways for an input compound are delineated based on a set of catabolic transformations (extracted from the literature and UM-BBD) "weighted" with experimental data on biodegradative fates extracted from databases (*see* **Note 2**) and with other factors such as the "biological oxygen demand."

There are many other alternatives, including commercial software. For a recent more exhaustive review, *see* [8]. Here we describe in detail the protocols for using two simple predictors of biodegradability which can be freely accessed through web interfaces. These two methods, described in detail above, represent two very different approaches for biodegradation prediction: an interactive, user-aided approach which gives information not only on the biodegradative fate of a compound but details on the biodegradative pathway(s) as well (UM-PPS), vs. a machine learning system (BDPServer/BiodegPred) which only predicts the final fate but can be applied to large collections of compounds since it is fully automatic.

2 Materials

The two resources described in the following protocols can be accessed through a standard web browser. Some of their features are implemented as *Java applets*. You might need to modify your browser's configuration and/or install some additional software for running these applets (*see* **Note 3**).

3 Methods

The two systems described here for the prediction of the biodegradative fate of chemical compounds are very simple to use. In the simplest case, entering the molecular structure of your target compound as single input and pressing a button is enough for obtaining the final result.

3.1 UM-PPS

1. The EAWAG-PPS (formerly UM-PPS) can be accessed at the following URL http://eawag-bbd.ethz.ch/predict/.
2. The only mandatory input for the system is the molecular structure of the chemical compound for which you want to predict the biodegradative fate (see below). Optionally, you can also specify the aerobic character of the environment in which the putative biodegradative process is going to take place (aerobic or anaerobic). If you know the SMILES string representation of your compound (*see* **Note 4**), enter it in the corresponding checkbox. If not, you can "draw" the molecular

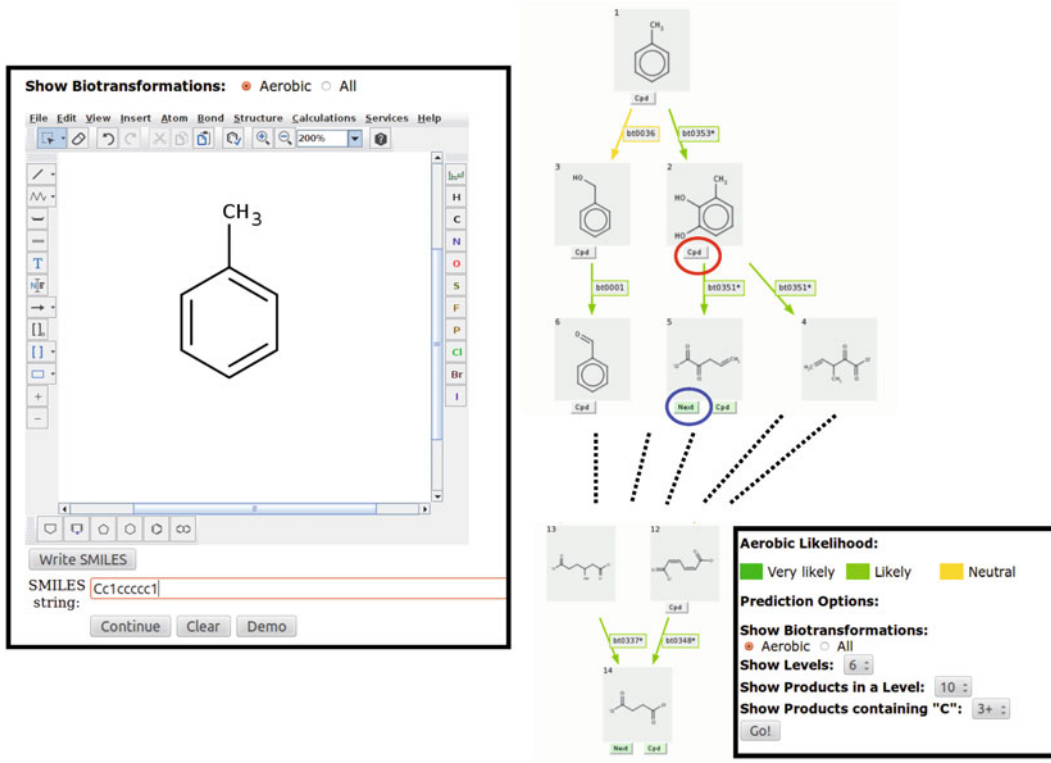

Fig. 1 Screenshots of the UM-PPS system when predicting the biodegradation routes for toluene (methylbenzene). The input form (*left*) includes a molecular editor to "draw" the structure of the input compound. In the predicted biodegradative routes (a portion of which is shown on the *right*), compounds present in the UM-BBD have a "Cpd" button (*red*) to go to the corresponding pages, while non-end compounds have a "Next" button (blue) to retrieve the downstream biodegradative routes starting with them

structure in the provided molecular editor and press "Write SMILES" afterward to automatically generate the SMILES string for the structure in the editor (Fig. 1).

3. Once the SMILES string of your compound is in place, press "Continue."
4. After a while, a representation of part of the predicted biodegradative pathway for your compound shows up (Fig. 1). In this representation, the aerobic likelihood of the different transformation steps is indicated by a color scale. The "rule" (transformation of functional groups) associated to each putative reaction is also shown (as its UM-BBD code, e.g., "bt0001"). These codes are active links to the UM-BBD pages with detailed information on the rules. Some of the compounds within this pathway (putative intermediate steps of the biodegradative process) might be present in the UM-BBD. In these cases, a "Cpd" label is included, which is an active link to the corresponding UM-BBD pages with detailed compound information.

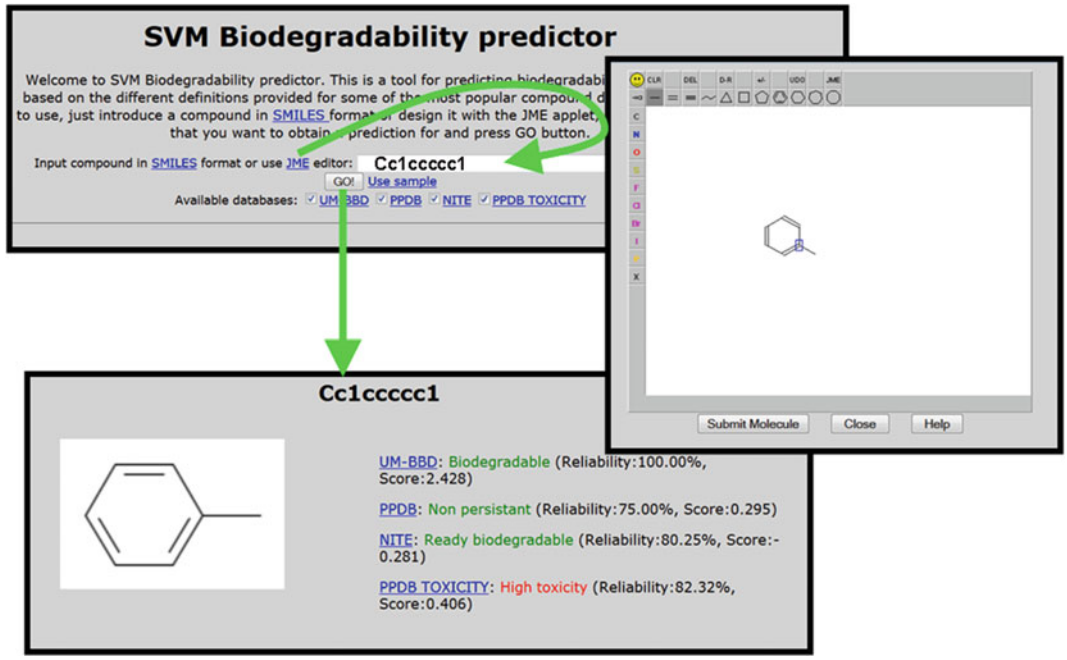

Fig. 2 Screenshots of the BiodegPred system when used for predicting the environmental fate of toluene (methylbenzene). The input form is at the *top* (including the molecular editor to enter the compound structure), and the results page is at the *bottom*

5. This initial representation does not contain all possible biodegradation routes generated by iteratively applying the rules. Only the first "n" biodegradative steps ("levels") and a given number of compounds per level are shown (see below). The "Next" labels below some of the compounds allow expanding the biodegradative routes starting at these compounds, allowing in this way to interactively explore the whole biodegradation network of your input compound. The idea is to select one route or another based on expert knowledge.
6. Finally, at the bottom of the page, a web form allows you to rerun the system for the same compound but changing the aerobic character or the number of levels and compounds per level shown.

3.2 BiodegPred

1. This system can be accessed at the following URL: http://csbg.cnb.csic.es/BiodegPred/.
2. As in the case of the UM-BBD, the only mandatory input for the system is the molecular structure of the compound. It can also be entered as a SMILES string (*see* **Note 4**) or drawn in the molecular editor following the "SME" link (Fig. 2). There is also a link ("Use sample") for filling the input textbox with an example structure.

3. As commented in the Introduction, this server predicts biodegradability (according with three different criteria) and toxicity. You can choose which of these four predictors you want to use with the provided checkboxes (all are selected by default).
4. To run the selected predictors on your input structure, press "Go."
5. The results page contains a representation of the chemical structure regenerated from the input SMILES (to check that it is correct) and the results of the predictors selected (Fig. 2). As explained earlier, this system only predicts the final "fate" of the compound and does not give information on the pathways used for reaching this final state. For each predictor, the results include the name of the database whose annotated compounds were used for training (UM-BBD, PPDB, NITE, and PPDB toxicity) which are active links to the corresponding resources. Next, you have the prediction for each database: "biodegradable" vs. "non-biodegradable" for UM-BBD (*see* **Note 5**), "persistent" vs. "non-persistent" for PPDB, "ready-biodegradable" vs. "non-ready biodegradable" for NITE, and "low toxicity" vs. "high toxicity" for PPDB toxicity. A color code is used for emphasizing the character of the predictions (green, biodegradable/nontoxic; red, recalcitrant/toxic). The predictor's score and the associated reliability are also indicated. The reliability values associated to each score were obtained from a test set of compounds of known fate, and they represent the fraction of compounds in the test set with that score or higher correctly predicted. Moving the mouse over these data, more information on the criteria used for defining these fates, on the scores, etc., is shown.
6. Note that the criteria used for classifying a compound as "biodegradable" or not in these three resources are different. For example, "persistent" (PPDB) is not exactly the same as "non-ready biodegradable" (NITE). Consequently the same compound could be annotated in different resources with apparently opposite fates, which translates also to the predictions. The user has to interpret these eventual apparent contradictions in view of the exact definition of the criteria.

4 Notes

1. The EPI Suite is available at http://www.epa.gov/opptintr/exposure/pubs/episuite.htm
2. There are many databases with different types of data related to microbial biodegradation of chemical compounds. These are not only useful for the developers of predictors (i.e., to retrieve

datasets for training/testing their systems) but also for the final user. That is because the biodegradation information of our compound of interest (or a similar one) might be already available in these resources. The University of Minnesota Biocatalysis/Biodegradation Database (UM-BBD), now EAWAG Biocatalysis/Biodegradation Database (EAWAG-BD) [11], is the main resource with information on known biodegradation routes (including data on compounds, reactions enzymes, microorganisms, etc.). The main database with general metabolic information, KEGG [15], "mirrors" the UM-BBD data on its "biodegradation of xenobiotics" pathways, so that this information can be queried and used in the same framework as the other KEGG pathways. There are also databases with experimental results on compound biodegradability, such as "half-lives" under different conditions, bioaccumulation, environmental toxicity, etc. For example, the Chemical Risk Information Platform (CHRIP) at the Japanese National Institute of Technology and Evaluation (NITE) (http://www.safe.nite.go.jp/english/) and the UK's Pesticide Properties Database (PPDB) (http://sitem.herts.ac.uk/aeru/projects/ppdb)

3. The molecular editors of the two resources commented are implemented as Java applets embedded in web pages. In recent versions of Java, in order for the embedded applets to work, you have to set the security level to "middle" in the Java configuration panel of your operative system. For example, in MS Windows: control panel > Java > security > security level > middle. Additionally, the first time the applet is run, you will have to accept a number of security warnings and "Allow...?" questions. You also need the *Java Runtime Environment* (JRE) installed on your system for applets to work (e.g., in MS Windows check whether a "Java" item is present in the control panel). You can download and install JRE from https://www.java.com/es/download/
4. The SMILE format (http://www.daylight.com) allows representing any chemical structure as a string of ASCII characters so that it can be stored and handled by computers. Most databases focused on chemical structures include the SMILE representation as a field. Consequently, if your compound is already stored in some database, you can copy/paste the SMILE string from there. If not, most software for converting among chemical formats allows converting from/to SMILES. Finally, there are a number of chemical editors online, such as those used in the two resources described here, which can generate SMILES for the structures entered by the user.
5. The "biodegradable"/"non-biodegradable" definitions for UM-BBD are based, as in the case of the BDPServer [10], on the possibility of finding a biodegradative pathway for the

compounds in the training set in this resource and are not internal annotations of the UM-BBD. Consequently, these annotations are themselves predictions and not experimental outcomes (as in the other three resources).

References

1. Diaz E (2004) Bacterial degradation of aromatic pollutants: a paradigm of metabolic versatility. Int Microbiol 7:173–180
2. Schmid A, Dordick JS, Hauer B, Kiener A, Wubbolts M, Witholt B (2001) Industrial biocatalysis today and tomorrow. Nature 409:258–268
3. Singh A, Ward OP (2004) Biodegradation and bioremediation. Springer, Berlin
4. Dua M, Singh A, Sethunathan N, Johri AK (2002) Biotechnology and bioremediation: successes and limitations. Appl Microbiol Biotechnol 59:143–152
5. Cases I, De Lorenzo V (2005) Genetically modified organisms for the environment: stories of success and failure and what we have learned from them. Int Microbiol 8(3):213–222
6. Williams ES, Panko J, Paustenbach DJ (2009) The European Union's REACH regulation: a review of its history and requirements. Crit Rev Toxicol 39(7):553–575
7. Wackett LP (2004) Prediction of microbial biodegradation. Environ Microbiol 6:313
8. Rücker C, Kümmerer K (2012) Modeling and predicting aquatic aerobic biodegradation - a review from a user's perspective. Green Chem 14:875–887
9. Kümmerer K (2007) Sustainable from the very beginning: rational design of molecules by life cycle engineering as an important approach for green pharmacy and green chemistry. Green Chem 9(8):899–907
10. Gomez MJ, Pazos F, Guijarro FJ, de Lorenzo V, Valencia A (2007) The environmental fate of organic pollutants through the global microbial metabolism. Mol Syst Biol 3:114
11. Gao J, Ellis LBM, Wackett LP (2010) The University of Minnesota biocatalysis/biodegradation database: improving public access. Nucleic Acids Res 38(D):D488–D491
12. Hou BK, Ellis LB, Wackett LP (2004) Encoding microbial metabolic logic: predicting biodegradation. J Ind Microbiol Biotechnol 31 (6):261–272
13. Wicker J, Fenner K, Ellis L, Wackett L, Kramer S (2010) Predicting biodegradation products and pathways: a hybrid knowledge- and machine learning-based approach. Bioinformatics 26(6):814–821
14. Oh M, Yamada T, Hattori M, Goto S, Kanehisa M (2007) Systematic analysis of enzyme-catalyzed reaction patterns and prediction of microbial biodegradation pathways. J Chem Inf Model 47(4):1702–1712
15. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32(Database issue):D277–D280
16. Dimitrov S, Kamenska V, Walker JD, Windle W, Purdy R, Lewis M, Mekenyan O (2004) Predicting the biodegradation products of perfluorinated chemicals using CATABOL. SAR QSAR Environ Res 15(1):69–82
17. Dimitrov S, Nedelcheva D, Dimitrova N, Mekenyan O (2010) Development of a biodegradation model for the prediction of metabolites in soil. Sci Total Environ 408 (18):3811–3816