D588–D592 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki068

MetaRouter: bioinformatics for bioremediation

Florencio Pazos*, David Guijas[1], Alfonso Valencia[2] and Victor De Lorenzo[2]

Department of Biological Sciences, Structural Bioinformatics Group, Biochemistry Building, Imperial College, London SW7 2AZ, UK, [1]bioALMA, Centro Empresarial Euronova, Ronda de Poniente, 4, 2nd floor, Unit C-D, 28760 Tres Cantos, Madrid, Spain and [2]National Center for Biotechnology, Cantoblanco, 28049 Madrid, Spain

Received August 14, 2004; Revised and Accepted October 5, 2004

ABSTRACT

Bioremediation, the exploitation of biological catalysts (mostly microorganisms) for removing pollutants from the environment, requires the integration of huge amounts of data from different sources. We have developed MetaRouter, a system for maintaining heterogeneous information related to bioremediation in a framework that allows its query, administration and mining (application of methods for extracting new knowledge). MetaRouter is an application intended for laboratories working in biodegradation and bioremediation, which need to maintain and consult public and private data, linked internally and with external databases, and to extract new information from it. Among the data-mining features is a program included for locating biodegradative pathways for chemical compounds according to a given set of constraints and requirements. The integration of biodegradation information with the corresponding protein and genome data provides a suitable framework for studying the global properties of the bioremediation network. The system can be accessed and administrated through a web interface. The full-featured system (except administration facilities) is freely available at http://pdg.cnb.uam.es/MetaRouter. Additional material: http://www.pdg.cnb.uam.es/biodeg_net/MetaRouter.

INTRODUCTION

Some microorganisms have acquired the ability to catabolize chemical compounds that do not form part of their central metabolism as they face them in the environment (1). This is being exploited for developing strategies aimed at the cleanup of pollutant compounds from soils and waters (bioremediation) (2).

Bioremediation offers many interesting possibilities from a bioinformatics point of view still slightly explored. This discipline requires the integration of huge amounts of data from various sources: chemical structure and reactivity of organic compounds; sequence, structure and function of proteins (enzymes); comparative genomics; environmental microbiology; and so on. The accumulation of huge amounts of data on individual genes and proteins allowed the first studies of biology from a 'Systems' perspective (3–7). From this point of view, biological systems are modeled as being composed of components in complex relationships whose ultimate properties cannot be understood by studying these components separately and later 'summing' their properties, but only by studying the system as a whole. In a similar manner, data related to bioremediation (genome sequences, structures of chemical compounds, enzyme sequences and structures, etc.) are being accumulated in public databases (8). This allows the first studies of bioremediation from a Systems Biology perspective (9,10), which complement the traditional approach focused on individual components (microorganisms, enzymes, etc.). The bioinformatics resources devoted to bioremediation are still scarce. Some interesting projects are being carried out to organize and store this huge amount of information related to this subject, The University of Minnesota Biocatalysis/Biodegradation Database (UMBBD) (8) being the more prominent resource.

Here we present MetaRouter, a system for maintaining heterogeneous information related to bioremediation and biodegradation in a framework that allows its updating, query, modification and mining. The core of the system is a relational database where the information on chemical compounds, reactions, enzymes and organisms is stored in an integrated framework. MetaRouter allows not only to interactively consult the database but also to formulate new questions with associated programs that run on top of the database.

THE DATABASE AND THE WEB INTERFACE

The current set of data include 740 chemical compounds (2167 synonyms), 820 reactions, 502 enzymes and 253 organisms.

*To whom correspondence should be addressed. Tel: +44 0 20 7594 5737; Fax: +44 0 20 7594 5789; Email: f.pazos@imperial.ac.uk

For the chemical compounds, the following information is included: name, synonyms, SMILES code, molecular weight, chemical formula, image of the chemical structure, canonical three-dimensional structure in PDB format, physicochemical properties (density, evaporation rate, melting point, boiling point and water solubility—the user can define and insert new ones) and links to other databases. For the reactions: substrates and products, catalyzing enzyme and links to other databases. For the enzymes: name, Enzyme Commission (EC) code, organisms where the gene is present, database sequence identifiers and links to other databases. The main public sources of information for obtaining the initial set of data were UMBBD (8), ENZYME (11) and Swiss-Prot (12).

The core database uses the PostgreSQL relational database management system. This database can directly be interrogated at the low level by the user using SQL queries or via data-mining programs. The system includes a web interface that allows to consult the database or to use the programs through standard web forms and to display the results in an interactive graphical way (Figure 1). A menu on the left-hand side of the interface, with the headings 'Reactions', 'Compounds' and 'Enzymes' provides a convenient way to start searching and browsing the database. It is possible to search for compounds, reactions or enzymes according to any of the properties described in the previous paragraph. Although it is not possible to make complex queries about the structure of chemical compounds, since the SMILES string contains a linear description of that structure, it can be used to perform some simple limited structural queries. In the representations of the results, all the elements (e.g. enzymes, reactions and chemical compounds) are active links that lead to other sections in the database or to external public resources. For example, in the pages containing information on chemical reactions, the substrates and products are linked to the corresponding pages with information on chemical compounds. Similarly, the enzymes are linked to the corresponding enzyme pages which in turn contain links to the entries for those enzymes in external sequence databases, like Swiss-Prot (12) and so on.

The client/server architecture allows the system to run in a central server and be used simultaneously by any number of client machines with the requirement of a standard web browser. For local/private installations of the system, a distinction between normal users and administrators (the ones who can modify the database) is incorporated. The administrator can add new information (for example, new reactions or compounds discovered/isolated in his/her laboratory) filling web forms in the web interface. Working with the system at a lower level (below the web interface) it is possible to incorporate new information on a large scale, for example, writing scripts to insert the full metabolism [KEGG (13)] into the SQL database.

A context-sensitive HTML-based help system is included.

DATA MINING

The current system for extracting new information from the database allows locating biodegradative pathways for an individual compound or a set of compounds. That is, it enables finding pathways between those compounds and the central metabolism. [We obtain the definition of 'central metabolism' from UMBBD (8). In general, it must correspond to all the pathways in KEGG (13) apart from 'biodegradation of xenobiotics'.] The system also permits to find pathways between one compound and another (or between two sets of compounds). The representation of possible pathways can be restricted according to a number of criteria: length of the pathway, all the required enzymes present in a given organism(s), all the intermediate compounds having a range of values for a given property (e.g. highly soluble), etc. The pathways are displayed together in a versatile representation where the user can choose what to see (compound names only or images of the chemical structures, synonyms, properties, formula, enzymes, etc.); and where the chemical compounds, reactions and enzymes can be colored according to their properties. All the elements in the representation (compounds, enzymes, reactions) are hyperlinked to the corresponding information pages in MetaRouter database. This system allows the exploration and design of biodegradative strategies for chemical compounds depending on conditions such as environment, bacterial ecosystem and others (see Figure 2 and Additional Material). Although a similar system for locating pathways is available at UMBBD (8), the one presented here provides graphical and more versatile representations and also additional features like the selection/restriction of pathways.

ADDITIONAL FEATURES

The system is designed in such a way that it is easy to incorporate other data-mining programs to perform more complex studies of the biodegradation network. This allows, among others, the characterization of the properties of the network from a Systems perspective (9), the generation of interactive representations of the whole network to visualize its global properties (see Additional Material) and the development of a system for predicting the biodegradative fate of chemical compounds based on their chemical structure that has been trained with the information present in the database (Gomez,M.J., manuscript in preparation). We intend to include access to many of these features in the web interface of the system.

DISCUSSION

The rise of genomic technologies and Systems Biology provide fresh approaches to currently untractable biological processes that are at the root of serious environmental problems. One formidable challenge in this respect is the biological fate of the nearly 8 million new chemical compounds (~40 000 predominant) which modern Organic and Industrial Chemistry has placed in the Biosphere. A large number of microbial strains are able to grow on environmental pollutants (about 800 today). Bioremediation was studied from a molecular biology point of view, characterizing the chemical reactions, genes, operons, etc. implicated in this process. The Biodegradation database of the University of Minnesota (8) has made a pioneering effort in putting together nearly every aspect of our current knowledge on biodegradation pathways and in developing systems for dealing with that data [e.g. to learn rules for predicting biodegradative features (14)]. Yet, most information available in the literature of microbial biodegradation of xenobiotics and recalcitrant chemicals deals with duos

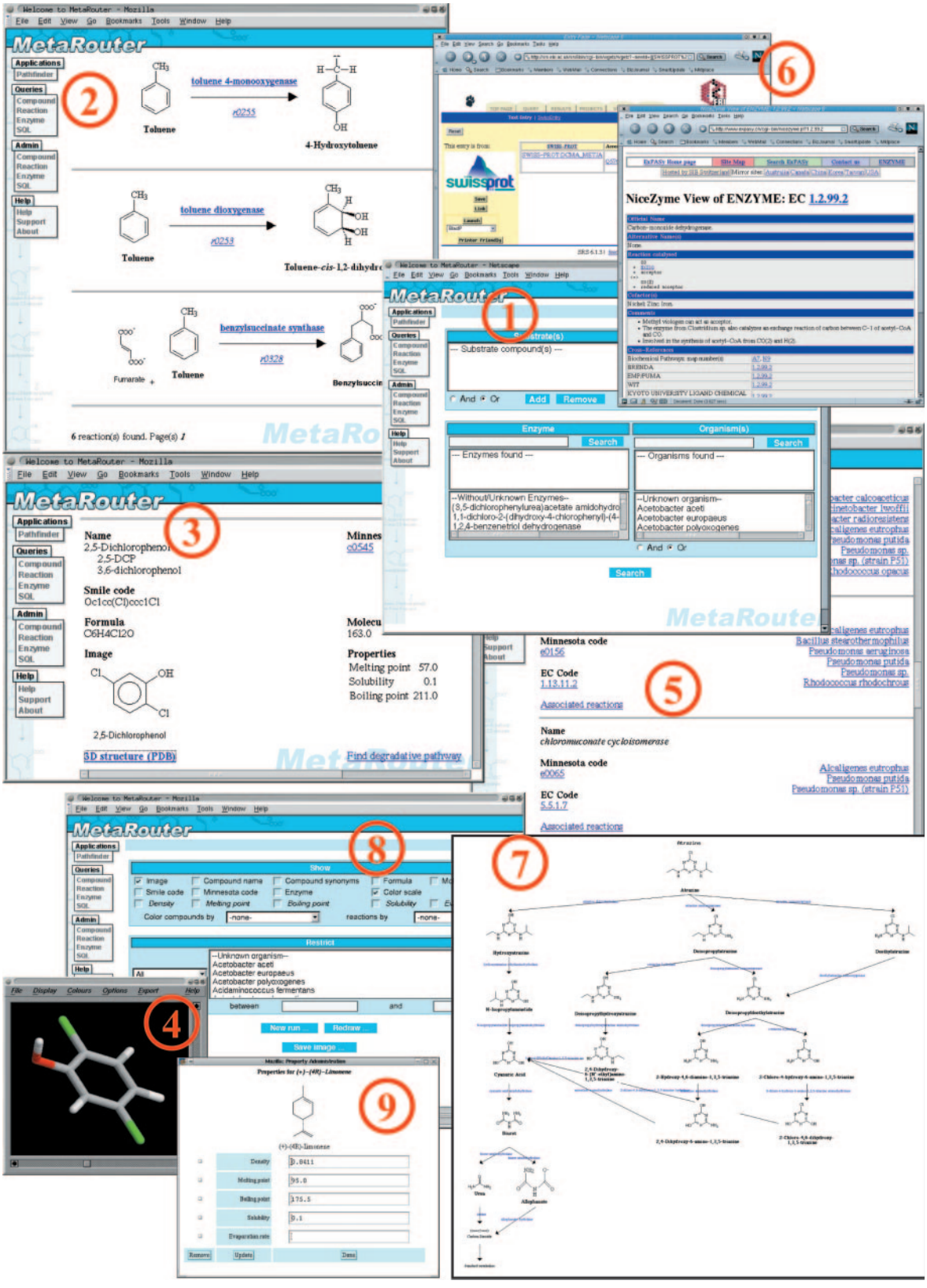

Figure 1. MetaRouter web interface. Some screenshots of a typical session using MetaRouter through its web interface are shown. All the dynamic HTML pages generated querying the database or using the system for locating pathways are linked between themselves and with external databases. (1) Searching for reactions; (2) reaction information; (3) chemical compound information; (4) chemical compound canonical 3D structure; (5) enzyme information; (6) linking to external databases; (7) possible pathways for the biodegradation of Atrazine; (8) controlling the representation of biodegradative pathways; and (9) editing the properties of a chemical compound.

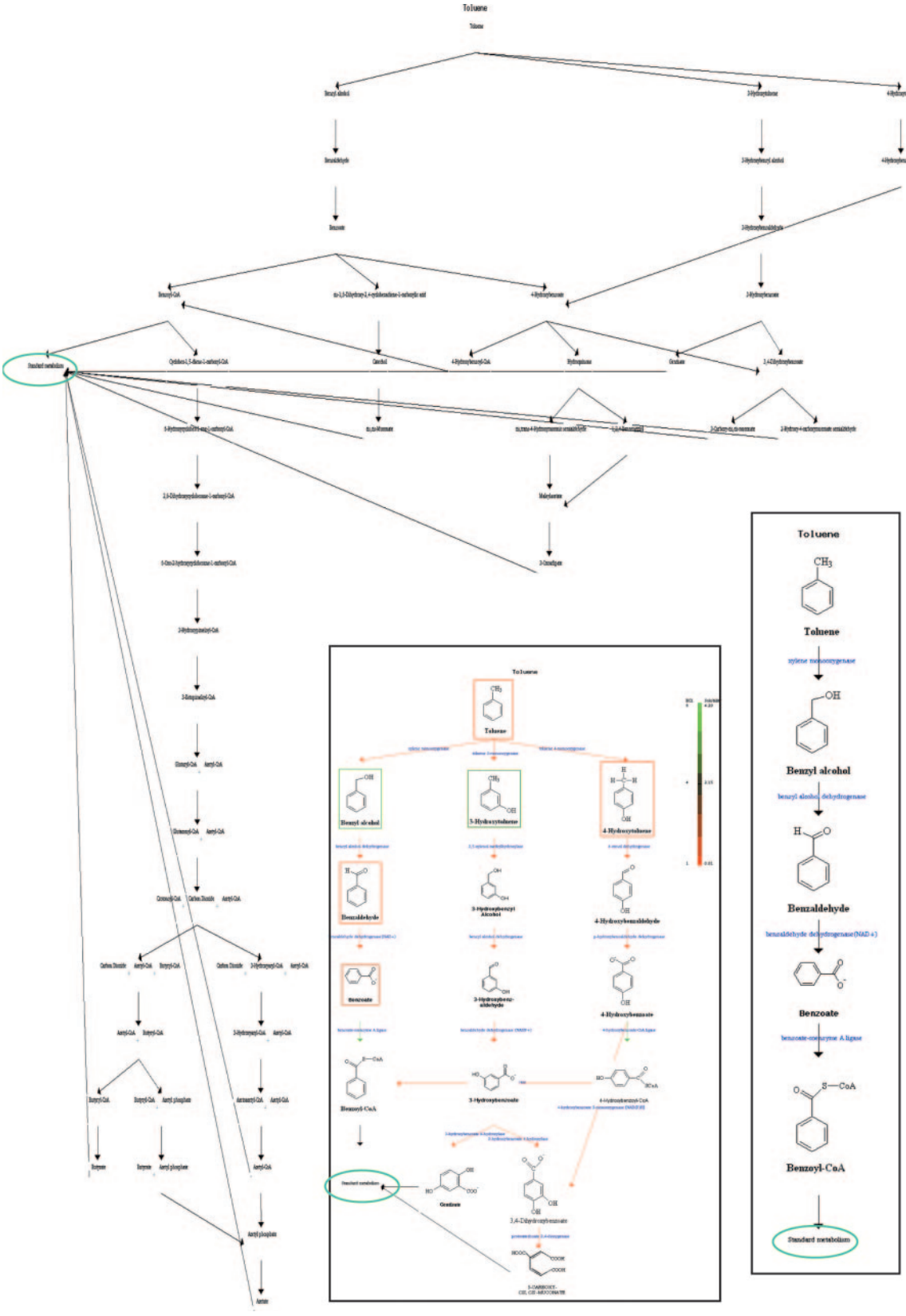

Figure 2. Biodegradative pathways for toluene (methyl-benzene) detected and represented with MetaRouter. The full network of possible pathways is represented displaying only the names of the chemical compounds. The subset of pathways that can be carried out by *Pseudomonas putida* is shown inside the central square displaying the enzymes, and coloring chemical compounds according to their water solubility and reactions according to the first digit of the EC code of the enzyme. The bottom-right square shows the shortest possible pathway for toluene degradation. The central metabolism is represented by a green circle.

consisting of one pollutant versus one strain and thus, lacks essential aspects of the natural scenarios, like the interchange of genes between bacteria (15) or their metabolic cooperation (16,17). This study of genomes and 'functionomes' from a community point of view (in contrast to organism point of view) is leading, for example, to the sequencing of 'genomes' of communities and ecosystems (18), instead of single organisms. These circumstances expose the need to qualify and to represent the information available in biodegradation databases in a fashion in which the entire known biodegradative potential of the microbial world can be crossed with the whole collection of compounds known to be partially or totally degraded through (mostly) bacterial action.

The database presented in this work constitutes a step towards the integration of heterogeneous sources of information related to bioremediation. Moreover, the associated querying and mining systems allow to extract new information from those data. The integration of heterogeneous data related to biodegradation from different sources was critical, for example, in proposing an evolutionary scenario for the biodegradation network (9). In that case, phylogenetic distributions of enzymes were compared with distances to the central metabolism. None of these two sets of data alone could have been used to address that problem.

The system presented here can help in assessing the environmental fate of compounds or mixtures and in designing biodegradative strategies for them. Among our future prospects are a plan to include in MetaRouter a system for predicting biodegradative pathways for new compounds not present in the database, the possibility of more advanced queries on chemical structures, links to MedLine and other databases, and an easy interface for studying and representing the global properties of the bioremediation network.

Available at the MetaRouter website, http://pdg.cnb.uam.es/biodeg_net/MetaRouter, are additional screenshots, more examples of biodegradative pathways, documentation of the system with a full description of its capabilities, examples of representations of the whole bioremediation network and its properties and information on how to access the system.

ACKNOWLEDGEMENTS

We want to acknowledge bioALMA's staff and Konrad Paszkiewicz (IC) for fruitful discussions, Jose María Fernandez (CNB) for technical support, and the members of Alfonso Valencia's laboratory (CNB), in particular those working in biodegradation (Manuel Gomez and Almudena Trigo) and Victor De Lorenzo's laboratory (CNB), especially Amalia Muñoz, for testing the program and giving feedback. Metarouter represents data derived, with permission, from The University of Minnesota Biocatalysis/Biodegradation Database (UM-BBD; http://umbbd.ahc.umn.edu/), obtained on May, 2002.

REFERENCES

1. Parales,R.E., Bruce,N.C., Schmid,A. and Wackett,L.P. (2002) Biodegradation, biotransformation, and biocatalysis (b3). *Appl. Environ. Microbiol.*, **68**, 4699–4709.
2. Dua,M., Singh,A., Sethunathan,N. and Johri,A.K. (2002) Biotechnology and bioremediation: successes and limitations. *Appl. Microbiol. Biotechnol.*, **59**, 143–152.
3. Ideker,T., Thorsson,V., Ranish,J.A., Christmas,R., Buhler,J., Eng,J.K., Bumgarner,R., Goodlett,D.R., Aebersold,R. and Hood,L. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science*, **292**, 929–934.
4. Jeong,H., Mason,S.P., Barabási,A.L. and Oltvai,Z.N. (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.
5. Alves,R., Chaleil,R.A.G. and Sternberg,M.J.E. (2002) Evolution of enzymes in metabolism: a network perspective. *J. Mol. Biol.*, **320**, 751–770.
6. Fraser,H.B., Hirsh,A.E., Steinmetz,L.M., Scharfe,C. and Feldman,M.W. (2002) Evolutionary rate in the protein interaction network. *Science*, **296**, 750–752.
7. Rison,S.G.C. and Thornton,J.M. (2002) Pathway evolution, structurally speaking. *Curr. Opin. Struct. Biol.*, **12**, 374–382.
8. 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.
9. Pazos,F., Valencia,A. and De Lorenzo,V. (2003) The organization of the Microbial Biodegradation Network from a Systems-Biology perspective. *EMBO Rep.*, **4**, 994–999.
10. Pieper,D.H., Martins dos Santos,V.A. and Golyshin,P.N. (2004) Genomic and mechanistic insights into the biodegradation of organic pollutants. *Curr. Opin. Biotechnol.*, **15**, 215–224.
11. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res.*, **28**, 304–305.
12. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
13. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.*, **32**, D277–D280.
14. Hou,B.K., Ellis,L.B. and Wackett,L.P. (2004) Encoding microbial metabolic logic: predicting biodegradation. *J. Ind. Microbiol. Biotechnol.*, **31**, 261–272.
15. Wilkins,B.M. (2002) Plasmid promiscuity: meeting the challenge of DNA immigration control. *Environ. Microbiol.*, **4**, 495–500.
16. Pelz,O., Tesar,M., Wittich,R.M., Moore,E.R., Timmis,K.N. and Abraham,W.R. (1999) Towards elucidation of microbial community metabolic pathways: unravelling the network of carbon sharing in a pollutant-degrading bacterial consortium by immunocapture and isotopic ratio mass spectrometry. *Environ. Microbiol.*, **1**, 167–174.
17. Abraham,W.R., Nogales,B., Golyshin,P.N., Pieper,D.H. and Timmis,K.N. (2002) Polychlorinated biphenyl-degrading microbial communities in soils and sediments. *Curr. Opin. Microbiol.*, **5**, 246–253.
18. Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D., Paulsen,I., Nelson,K.E., Nelson,W. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.