

Bionemo: molecular information on biodegradation metabolism

Guillermo Carbajosa, Almudena Trigo, Alfonso Valencia and Idefonso Cases*

Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Melchor Fernández Almagro, 3, E-28029, Madrid, Spain

Received October 1, 2008; Revised October 16, 2008; Accepted October 16, 2008

ABSTRACT

Bionemo (<http://bionemo.bioinfo.cnio.es>) stores manually curated information about proteins and genes directly implicated in the Biodegradation metabolism. When possible, the database includes information on sequence, domains and structures for proteins; and sequence, regulatory elements and transcription units for genes. Thus, Bionemo is a unique resource that complements other biodegradation databases such as the University of Minnesota Biocatalysis/Biodegradation Database, or Metarouter, which focus more on the biochemical aspects of biodegradation than in the nature of the biomolecules carrying out the reactions. Bionemo has been built by manually associating sequences database entries to biodegradation reactions, using the information extracted from published articles. Information on transcription units and their regulation was also extracted from the literature for biodegradation genes, and linked to the underlying biochemical network. In its current version, Bionemo contains sequence information for 324 reactions and transcription regulation information for more than 100 promoters and 100 transcription factors. The information in the Bionemo database is available via a web server and the full database is also downloadable as a PostgreSQL dump. To facilitate the programmatic use of the information contained in the database, an object-oriented Perl API is also provided.

INTRODUCTION

The cleanup of pollutants from soil and water is becoming an important task as human activities produce large amounts of chemical compounds (up to tens of millions)

that are frequently released into the environment. Microbial populations play an important role in this process as they have acquired the ability to metabolise these compounds using them as carbon and energy sources (1). Decades of biochemical studies have produced a considerable wealth of knowledge about this unique metabolism, and this has recently started to be categorized and stored in structured databases. Examples of these are the University of Minnesota Biocatalysis/Biodegradation Database (UM-BBD) (2) and The Metarouter (3). Although this formalization has permitted the development of useful applications, such as a predictor of chemical biodegradability (4,5), the absence of information at the sequence level of the proteins required for the biodegradation process has limited further systematic studies. Questions such as the molecular basis of enzyme specificity, their catalytic mechanism, the evolutionary origin of this novel metabolism, or the spreading of such activities in the environment, are extremely difficult to address in the absence of accurate sequence and genetic information.

Current biodegradation databases such as UM-BBD or Metarouter link reactions to protein sequences in databases that have been annotated with the corresponding Enzyme Commission Code (EC Code). Although we appreciate its value, this method can be inaccurate. For instance, many reactions share the same EC Code although they use distinct substrates and generate different products. For example, the EC code 1.13.11.- is shared by more than 40 reactions, including a large number of dioxigenases with different substrates such as styrene or pyrene. Given that the Uniprot Database contains more than 200 sequences annotated with this particular EC code (1.13.11.-), it is impossible to use the information provided in the biodegradation databases to accurately associate the protein that actually carries out a specific biochemical reaction. In some cases, the databases provide links to The Kyoto Encyclopedia of Genes and Genomes (KEGG), but unfortunately KEGG associations between proteins and reactions, are often inaccurate, and direct to

*To whom correspondence should be addressed. Tel: +34 917 328 000; Fax: +34 912 246 980; Email: icases@cnio.es

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

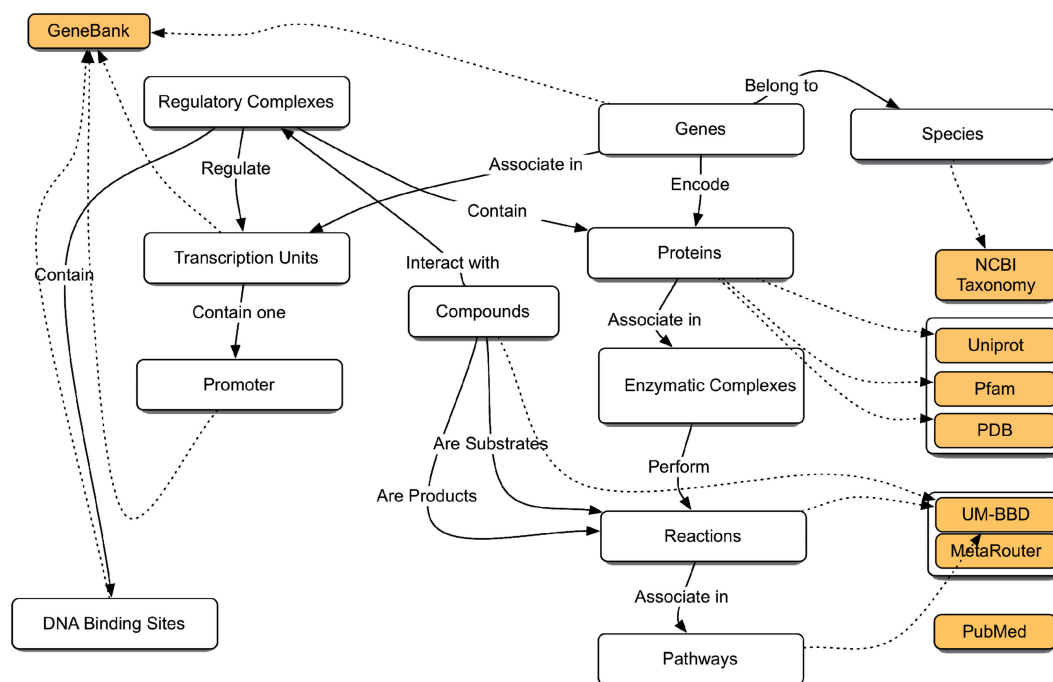


Figure 1. Schematic representation of Bionemo data entities and their relations. All the biological entities contained in Bionemo are shown as white boxes, with black arrows indicating their relationships. Entities are sorted in two vertical axes, the one on the left showing entities related to regulation and the one on the right showing entities related to metabolism. Both classes are linked by chemical compounds. Orange boxes are external databases to which the different entities are connected, as indicated by dashed arrows.

proteins electronically annotated in the course of complete genome sequencing projects. As an alternative method, UM-BBD allows searches with the enzymatic activity name as query directly in GenBank, but usually it is difficult to identify if the association between the protein and the biochemical activity is based on a published experimental characterization or whether it was inferred by computational methods.

As an alternative to these approximations, Bionemo provides accurate associations between proteins and reactions based on customized database searches, extensive literature mining and manual curation. Bionemo offers an additional level of information by adding molecular data on the transcriptional organization of biodegradation genes and their regulation. This molecular information is placed in the proper metabolic context and thus enables the exploration of particular features that have traditionally been difficult and tedious to address. In summary, Bionemo complements the currently available information on biodegradation that focus in the biochemical aspects by adding additional layers of molecular information at the level of proteins and gene control.

DATABASE CONTENT AND DATA ACQUISITION

The Bionemo Database combines metabolic, genetic and regulatory information. The central entities of the database are the enzymatic complexes. These are linked to the biochemical reactions defined as the transformation of substrates into products. Reactions associate into pathways. Enzymatic complexes are composed of protein

subunits that are encoded by genes. Genes are associated into transcription units that contain a promoter. These transcription units are regulated by one or more regulatory complex, comprising of a transcription factor, one or more DNA-binding sites and generally an inducer compound. All entities in the database are linked to the corresponding external databases, such as GenBank or The NCBI Taxonomy Database (6), Uniprot (7), Pfam (8) and UM-BBD (2) (Figure 1).

The first step in the manual curation process of Bionemo consisted of obtaining all the pathways and reactions from the University of Minnesota Biocatalysis/Biodegradation Database (July 2005). For each reaction, we manually searched sequence databases, EMBL (9) and GenBank (6) using as many query terms as was possible, including EC number, reaction name, enzyme name or original publication as reported in the UM-BBD. The resulting entries were then screened manually and the sequences were associated to a reaction only if we could find a journal article describing that particular enzymatic activity for the gene using the same name as in the sequence database, and that refers to the same organism, including the strain name. It is important to stress that all the connections between protein complexes and biochemical reactions were established manually and were always based on the information contained in the scientific literature. No sequence similarity information or other computational function prediction methods were used, nor were articles using those methods considered as evidence. In total, we obtained accurate and reliable protein associations for 324 reactions of the original set of 945 (Table 1). A similar process was followed to retrieve genes and

Table 1. Summary of the Bionemo information content

Pathways	145
Reactions	945
with associated complex	324
Enzymatic complexes	537
Proteins	1107
Microbial species	234
Transcription units	212
Transcription factors	90
Effectors	90
TF DNA-binding sites	128
Promoters	100

transcriptional units. In the case of regulatory proteins, both their binding sites and the regulatory actions on regulated promoters were also obtained from the literature, starting from the reviews by Tropel and van der Meer (10) and Diaz and Prieto (11), and following references therein.

ACCESS AND DISTRIBUTION

The main way to access Bionemo is via its web site (<http://bionemo.bioinfo.cnio.es>). While a full list of each entity type is provided to browse the Bionemo content, the web server implements a simple search interface that allows simultaneously querying all the biological entities described above. The results are shown categorized by tabs representing classes containing the entity types (reactions, complexes, etc.). From the results page, the user can easily access entity-specific pages, in which all information available is summarized, including the biochemical, sequence and regulatory data. Links to external databases are provided, including the original UM-BBD metabolic information, GenBank and Uniprot, the NCBI Taxonomy database for microbial species, and the PubMed references to the original information sources.

Navigating the database is simple. To illustrate this, we could start with the search for the chemical compound 'benzoate'. The results page returns all the available entities in Bionemo with a partial match to that query, categorized in four classes: 54 reactions, 22 pathways, 22 enzymatic complexes and 24 compounds. Without navigating away from this page, the user can, for instance, access all the protein complexes that perform reactions related to benzoate or any derivative, and the gene that encode them. The user can also access the reactions in which any of the 24 compounds take part, either as product or substrate, and the 22 pathways they belong to. Should the user select the benzoate pathway (Benz2), a graphical representation is shown (Figure 2) that is also conveniently clickable. Transcriptional regulation, when available, can be shown on top of the pathway if required. As always, a link to the original source in UM-BBD and connections to alternative pathways are also shown in this page, and we have made clickable all the compounds, reactions and regulators included in the graph. For instance, selecting the halobenzoate 1–2 dioxygenase link will take the user to the page for that reaction. This page

includes a link to UM-BBD for a full description of the reaction. Bionemo provides the list of all complexes capable to perform that reaction along with links to the articles we used to establish that association, and links to additional relevant articles describing either the reaction or the enzymatic complexes. The user can now select a gene, *cbdA* from *Burkholderia* sp. TH2 in the example shown, and access its page. There, links to several external databases are offered, including GenBank for DNA sequences and GenePep and/or Uniprot for the protein sequences. In addition, if the protein sequence contains domains included in Pfam, these are shown and properly connected to Pfam. Clicking the domain name will retrieve the list of all proteins in Bionemo containing that domain. Finally, if available, links to resolved 3D structures in the PDB are also provided.

Regarding the transcriptional information, the user can continue the exploration of the system by using the corresponding link. In this case, the gene *cbdA* contains information about transcriptional regulation indicated by the transcriptional unit (the *cbdABC* operon) and by the transcription factor that regulates the gene: CbdS (12). The transcription unit page contains a scaled schematic representation of the operon and its promoter region, including the coordinates of the transcription factor binding sites if known, the type of promoter that drives its expression, the type of regulation (activation or repression) and the molecules that can serve as effectors of the system. In the CbdS transcription factor page, in addition to the standard gene information (links to external sequence databases, domain structure and relevant articles), the list of target transcription units and the molecules that act as effectors are also included.

As seen in this example, the Bionemo web site has been designed to facilitate browsing and fast access to relevant information. Bionemo has been intentionally designed to avoid redundancy with related databases, and we always provide external links to all of them. For advanced users, Bionemo can be downloaded as a SQL dump and installed locally. A Perl API (application program interface) is also provided in order to facilitate programmatic access and to avoid the use of complex SQL queries. The local installation process and the use of the API are properly documented in the Bionemo web site.

APPLICATIONS AND FUTURE DIRECTIONS

Bionemo has a range of application in the fields of molecular biology and genomics. Information contained in Bionemo can be useful for cloning, primer design for PCR amplification and design of directed mutations, among other applications. For the annotation of genomes and metagenomic libraries, Bionemo is a better-suited alternative to the non-specialized sequence databases. In the field of sequence analysis, Bionemo allows the generation of sequence alignments enriched in experimentally characterized proteins, which can be useful for finding residues related with substrate specificity or catalytic mechanism. Beyond these applications, we envision a wide range of emerging fields in which Bionemo can be useful,

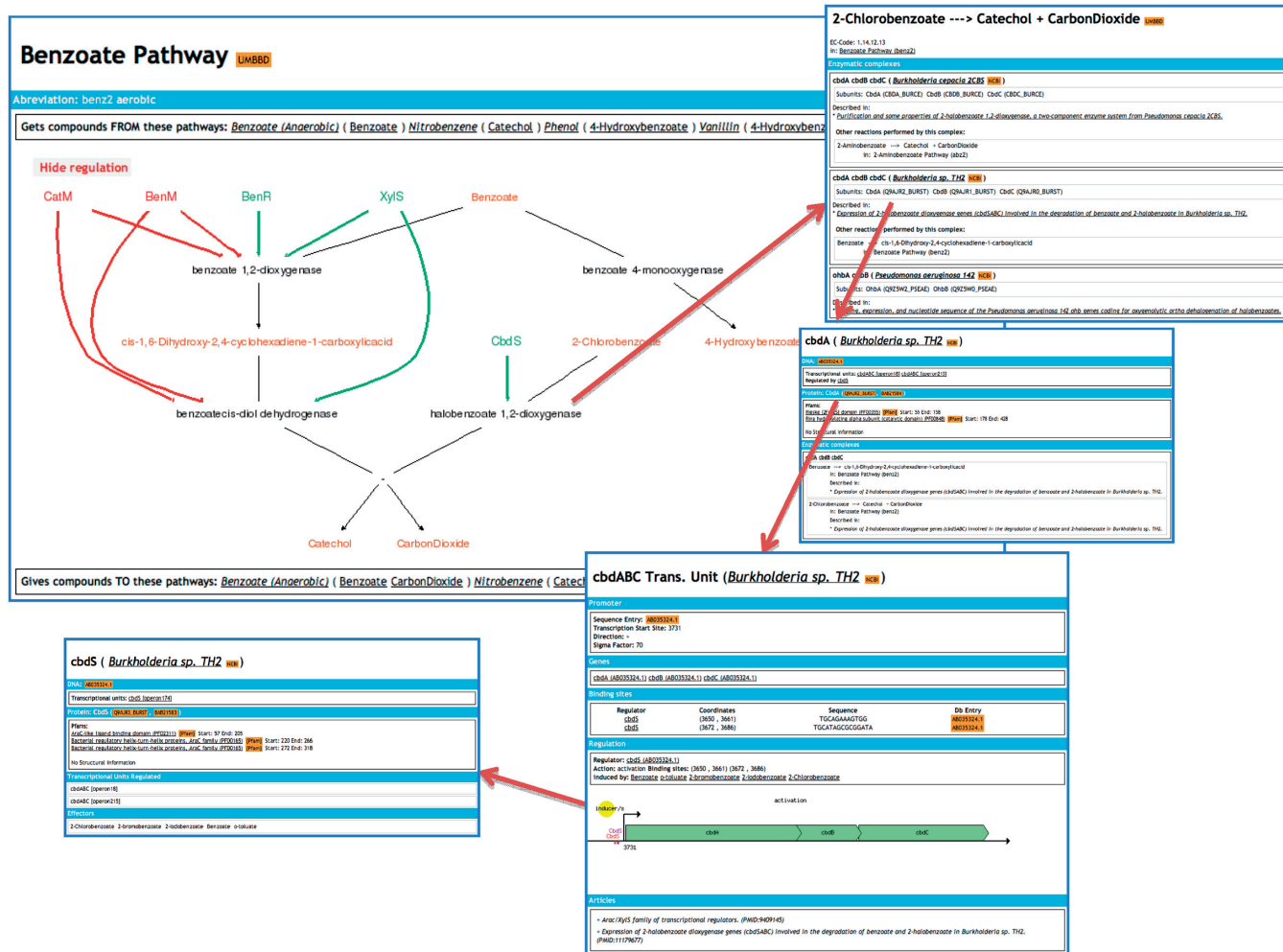


Figure 2. Example of information retrieval using the Bionemo web site. A standard case of use could start by browsing a pathway and its regulatory elements. From there, the user can navigate to different views. Then, by clicking on a reaction name, the user will be taken to a new page showing all the enzymatic complexes able to perform that reaction, and from there to one describing proteins, their coding genes, their transcriptional organization and the transcription factors involved.

such as systems biology of biodegradation. This field has started to produce relevant results (13) and can be enhanced by the availability of protein sequences and regulatory information. This information will allow researchers to dive deeper in the functional properties of the enzymes and their particular organization. But proteins are only a small part of the whole picture, where transcriptional events are the driving forces of evolution. Therefore, making accessible the genetic organization of biodegradation enzymes will allow a better characterization of their ability to be transferred among different organisms, a process that has been reported as key in the biogeochemical cycles of toxic compounds (14,15). From an engineering point of view, Bionemo can also help in the designing of new pathways and regulatory circuits, in which sequence information and protein-protein and protein-DNA interactions are required for the proper design of useful and independent systems (16-18). In this context, Bionemo will be useful to assist the design of new transcription factors

with the desired specificity and with a predictable behaviour (19).

In order to expand the reaction coverage of Bionemo, we are currently implementing novel tools that will help us keep the result up to date. Our laboratory is currently working on a set of text mining tools that will scan the literature for precise biodegradation information that will be added to Bionemo under expert curation. These tools will be able to mine relevant information: not only the database entries associations but also new biochemical reactions, the species involved and particular environment in which they take place.

To summarize, Bionemo is a unique resource that contains several layers of integrated information and enables rational and comprehensive access to the biochemical pathways, protein complexes, and genetic regulation of biodegradation. To the best of our knowledge this is the only resource available dealing with all molecular aspects of biodegradation. We therefore expect this tool to be used in a broad spectrum of scientific and applied research.

ACKNOWLEDGEMENTS

We thank Michael Tress for reviewing the manuscript.

FUNDING

This work is funded by the EMERGENCE EU grants, the pSYSMO project within the SYSMO Framework, and the Fundación Banco Bilbao Vizcaya Argentaria (FBBVA-BIOCON-3). I.C. is a member of the Ramón y Cajal Program of the Spanish Ministry of Education and Science. Funding for open access charge: European Union.

Conflict of interest statement. None declared.

REFERENCES

1. Diaz,E. (2004) Bacterial degradation of aromatic pollutants: a paradigm of metabolic versatility. *Int. Microbiol.*, **7**, 173–180.
2. Ellis,L.B., Roe,D. and Wackett,L.P. (2006) The University of Minnesota Biocatalysis/Biodegradation Database: the first decade. *Nucleic Acids Res.*, **34**, D517–D521.
3. Pazos,F., Guijas,D., Valencia,A. and De Lorenzo,V. (2005) MetaRouter: bioinformatics for bioremediation. *Nucleic Acids Res.*, **33**, D588–D592.
4. Gomez,M.J., Pazos,F., Guijarro,F.J., de Lorenzo,V. and Valencia,A. (2007) The environmental fate of organic pollutants through the global microbial metabolism. *Mol. Syst. Biol.*, **3**, 114.
5. Ellis,L.B., 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.
6. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2007) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **35**, D5–D12.
7. Boutet,E., Lieberherr,D., Tognolli,M., Schneider,M. and Bairoch,A. (2007) UniProtKB/Swiss-Prot: the manually annotated section of the UniProt KnowledgeBase. *Methods Mol. Biol.*, **406**, 89–112.
8. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al.* (2007) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
9. Kulikova,T., Akhtar,R., Aldebert,P., Althorpe,N., Andersson,M., Baldwin,A., Bates,K., Bhattacharyya,S., Bower,L., Browne,P. *et al.* (2007) EMBL Nucleotide Sequence Database in 2006. *Nucleic Acids Res.*, **35**, D16–D20.
10. Tropel,D. and van der Meer,J.R. (2004) Bacterial transcriptional regulators for degradation pathways of aromatic compounds. *Microbiol. Mol. Biol. Rev.*, **68**, 474–500.
11. Diaz,E. and Prieto,M.A. (2000) Bacterial promoters triggering biodegradation of aromatic pollutants. *Curr. Opin. Biotechnol.*, **11**, 467–475.
12. Suzuki,K., Ogawa,N. and Miyashita,K. (2001) Expression of 2-halobenzoate dioxygenase genes (cbdSABC) involved in the degradation of benzoate and 2-halobenzoate in *Burkholderia* sp. TH2. *Gene*, **262**, 1374–15.
13. 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.
14. 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.
15. 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.
16. Cases,I. and 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**, 213–222.
17. Cases,I. and de Lorenzo,V. (2005) Promoters in the environment: transcriptional regulation in its natural context. *Nat. Rev. Microbiol.*, **3**, 105–118.
18. Silva-Rocha,R. and de Lorenzo,V. (2008) Mining logic gates in prokaryotic transcriptional regulation networks. *FEBS Lett.*, **582**, 1237–1244.
19. Galvao,T.C., Mencia,M. and de Lorenzo,V. (2007) Emergence of novel functions in transcriptional regulators by regression to stem protein types. *Mol. Microbiol.*, **65**, 907–919.