JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

ARTICLE

pubs.acs.org/jcim

A Simple Protocol for the Comparative Analysis of the Structure and Occurrence of Biochemical Pathways Across Superkingdoms

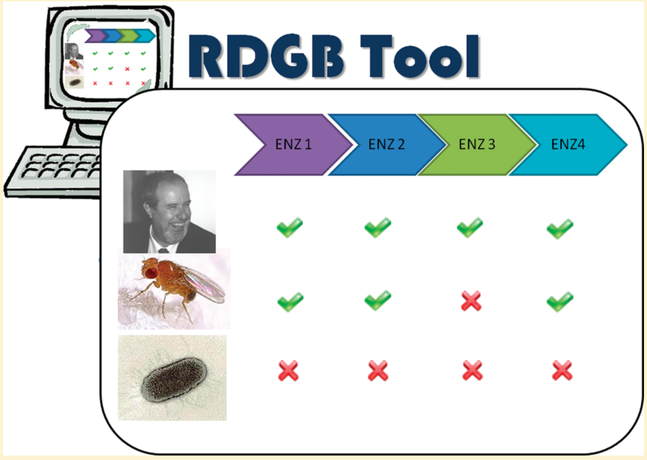
Claudia Andreini,[†,‡] Ivano Bertini,*[,†,‡] Gabriele Cavallaro,[†] Leonardo Decaria,[†] and Antonio Rosato[†,‡]

[†]Magnetic Resonance Center (CERM), University of Florence, Via L. Sacconi 6, 50019 Sesto Fiorentino, Italy

[‡]Department of Chemistry, University of Florence, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy

Supporting Information

ABSTRACT: A biochemical pathway can be viewed as a series of chemical reactions occurring within a cell, each of which is carried out by one or more biological macromolecules (protein, RNA, or complexes thereof). Computational methods can be applied to assess whether one organism is able to perform a biochemical process of interest by checking whether its genome encodes all the components that are known to be necessary for the task. Here we present a simple strategy for collecting the above data that is based on, but not limited to, our experience on processes involving metal ions and metal-binding cofactors. The strategy is fully implemented in a bioinformatics package, Retrieval of Domains and Genome Browsing (RDGB), which is available from http://www.cerm.unifi.it/home/research/genomebrowsing.html. The use of RDGB allows users to perform all the operations that are needed to implement the aforementioned strategy with minimal intervention and to gather all results in an ordered manner, with a tabular summary. This minimizes the (bio)informatics needed, thus facilitating its use by nonexperts. As examples, we analyzed the pathways for the degradation of organic compounds containing one or two aromatic rings as well as the distribution of some proteins involved in Cu_A assembly in more than a thousand prokaryotes.



■ INTRODUCTION

Thanks to the success of genome sequencing projects, it is possible to perform experimental and computational studies at the whole genome and/or whole proteome level, which leverage on the availability of a potentially complete list of the proteins codified by a living organism. In particular, there has been a great deal of interest in the identification of so-called functional modules, i.e., groups of proteins working together for the same cellular function. A typical case is that of enzymatic pathways in metabolic networks. Over the years, a huge portfolio of tools has been developed by a great number of different bioinformaticians all over the world to reconstruct such modules by computationally predicting functional relationships among the proteins encoded by genomes. Three main features have been exploited to this aim: the occurrence of gene fusion events (Rosettastone);[1,2] the conservation of gene order;[3−5] and the similarity of phylogenetic profiles.[6,7] Combinations of these (and other) methodologies have also been developed.[8−10] The intended use of the results output by these tools, which typically include the identity of the partners of each protein along a pathway in one or more organisms and their functional linkages, is to drive experimental studies aimed at, e.g., defining the role of uncharacterized proteins in the pathway[11−13] or supporting more complex computational studies.[14−16] Alternatively, it is possible to obtain information on already characterized pathways thanks to specialized databases, such as KEGG,[17,18] BioCyc,[19] or Reactome.[20] These databases contain information on metabolic pathways extracted from the relevant literature with manual curation. Within the frame described here, they are useful to identify which enzymes are known to be part of a given pathway.

Despite this wealth of available resources, it is not entirely obvious how to use the knowledge of the protein components that make up a biochemical pathway in an individual organism or set of organisms for tasks like the investigation of how widespread is a process throughout the domains of life or the identification of its possible variants. Whereas there are again many computational tools available to facilitate these tasks, setting up a consistent strategy to reliably use them against hundreds of genomes is not trivial. We have extensively faced this difficulty when trying to investigate pathways in the biosynthesis, assembly, or transport of different metal-containing

Received: October 7, 2010
Published: March 10, 2011

cofactors[21,22] or to characterize the occurrence in proteomes of some metalloproteins.[21−25] To tackle this kind of study, we have developed a number of scripts and programs that, e.g., automate the download and interrogation of databases or the identification of specific amino acidic patterns, such as metal-binding patterns.[26] Commonly, the proteome-level analysis of the occurrence of a biochemical pathway is based on the identification of homologues of all the involved proteins. Again, various computational methods can be used to this end. The detection of bidirectional best hits, often done with the BLAST program,[27] is one of the most widely used approaches.[3,28] Another approach is based on the identification of conserved domains through the use of profiles.[29]

In this work, we describe a coherent, easy protocol for the identification of a set of proteins that can constitute an entire biochemical pathway on the basis of homology relationships detected through the presence of conserved domains and integrating, when available, 3D structural information. This protocol integrates all the tools that we developed and tested in our previous publications[22] into a single package, which we called RDGB (Retrieval of Domains and Genome Browsing). RDGB not only integrates all the needed scripts and makes them easy to use for nonexperts but also enforces the use of a tested, internally consistent protocol in order to guarantee the reliability of the results. In addition, it provides a preordered manner of storing the data which can be useful for subsequent analyses as well as further computational analyses.

As an example, we analyzed the aerobic degradation of aromatic hydrocarbons in 1136 completely sequenced prokaryotic genomes. Aromatic hydrocarbons, such as toluene or biphenyl, are common contaminants of soil and groundwater and are listed as priority pollutants by the U.S. Environmental Protection Agency,[30] either as single compounds or in mixtures. One of the most attractive means to remove these compounds from polluted environments is through bioremediation.[31,32] Numerous bacterial strains have been isolated for the ability to aerobically degrade a variety of aromatic hydrocarbons.[33] The genes encoding the enzymes needed for the biodegradation of aromatic hydrocarbons can be located either in plasmids or in chromosomal DNA. The bacterial degradation of aromatic hydrocarbons consists of many reaction steps, which have often been broadly separated into peripheral and central pathways. Peripheral pathways convert a large proportion of different aromatic hydrocarbons into a limited number of key central intermediates, such as catechol and protocatechuate. The aerobic degradation of aromatic compounds is frequently initiated by ring-hydroxylating oxygenases,[34] which catalyze the incorporation of two oxygen atoms into the aromatic ring to form arene cis-diols, followed by a dehydrogenation reaction catalyzed by a cis-dihydrodiol dehydrogenase to give catechol or substituted catechols which serve as substrates for oxygenolytic aromatic ring cleavage.[35]

As a further example, we investigated the occurrence of proteins involved in the assembly of the Cu_A cofactor. Cu_A is a redox-active cofactor that contains two copper ions; in the reduced state both ions are in the +1 state. Upon one-electron oxidation of the cofactor, a mixed-valence species forms where the two copper ions are formally in the +1.5 oxidation state. The Cu_A cofactor is contained within the soluble domain of subunit II (Cox2) of prokaryotic and eukaryotic cytochrome *c* oxidases or within a homologous C-terminal domain of prokaryotic nitrous oxide reductase. Its physiological function is to shuttle the electron that it acquires from cytochrome *c* (either soluble or membrane anchored) to the catalytic core of the enzyme, where it is used to reduce dioxygen or nitrous oxide, respectively. The correct assembly of Cu_A is crucial for enzyme's function. The assembly process is relatively complex, and a number of ancillary proteins have been implicated in it.[36] NMR studies have demonstrated that in *Thermus thermophilus* copper(I) ions are delivered to the Cu_A binding site of Cox2 by a periplasmic metallochaperone called PCu_AC, while a second protein, Sco1, is responsible for maintaining the correct oxidation state of the Cys ligands in the Cu_A binding site by acting as a thiol-disulfide oxidoreductase.[37] Interestingly, Sco1 can also bind copper(I) or copper(II) ions, but this ability does not seem important for the assembly of Cu_A; an interplay between the oxidoreductase and metallochaperone activities of Sco1 proteins has been proposed based on computational studies.[24]

METHODS

Overview of the Computational Approach. The RDGB tool can be run on computers having Linux as their operating system. It is written in Python and uses a variety of different scripts and programs, which we have developed in the past few years,[21−25] contained in the subfolder *0Tools* that is created upon installation. From the user's point of view, it is important to note that RDGB is divided in two main Python scripts: *Retrieving_domains.py* and *Genome_browsing.py*, which are described in detail below. The two scripts are run consecutively as the first one builds part of the input to the second script. Python version 2.4.3 or higher is needed, with the following modules installed: *ftplib*, *os*, *re*, *string*, *time*, and *urllib*. All these modules are in the standard Python release.[38]

In the present strategy, we use the protein domains defined in the Pfam library[39,40] to identify putative homologues of the proteins involved in the pathway in any desired genome or list of genomes. When not already known, the domains can be initially identified in the sequence of proteins of known 3D structure that are available from the Protein Data Bank (PDB).[41] In our experience this is quite useful when trying to collect ensembles of proteins that can bind the same ligand, as sometimes not all the domains that can do this have been annotated as such in Pfam. Instead, if the ligand is present in the 3D structure of the protein, this information can be readily extracted from the PDB database together with the pattern of amino acids that are involved in the interaction of the protein with it. The latter is called the ligand binding pattern (LBP) and is defined by the identity and spacing of the amino acids, e.g., $CX_4CX_{20}H$, where X is any amino acid. As discussed in more detail in the next sections, this pattern can be usefully applied as a filter to reduce the number of false positives (i.e., of the proteins predicted to bind the cofactor but which in reality are unable to bind it) by rejecting the proteins that lack the LBP. The script *Retrieving_domains.py* performs the identification of domains in PDB structures and of the corresponding LBP's, and downloads the relevant hidden Markov models (HMM's)[42] that describe the domains from the Pfam database for the subsequent proteome searches. These data can be used independently.

The *Genome_browsing.py* script, which should be run after *Retrieving_domains.py*, downloads the proteomes of the organisms of interest from the National Center for Biotechnology Information (NCBI) database[43] and then identifies the sequences containing the domains previously retrieved executing *Retrieving_domains.py*. For the latter step it uses the HMMER 3.0

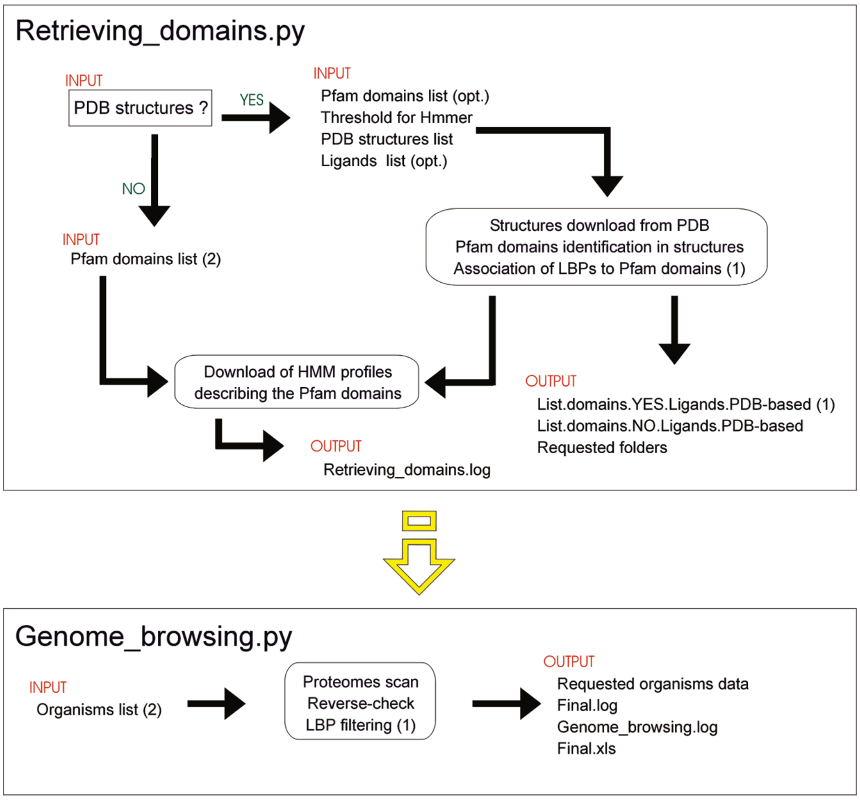


Figure 1. Flowchart describing the RDGB tool. The tool is composed by two main scripts, which should be run sequentially. The first script, *Retrieving_domains.py* (top box), accepts as input a list of Pfam domains and/or a list of PDB structures. If the latter is provided, then the user can additionally input a list of ligands, which will be used to define LBP's (see text). The list of PDB structures is used to identify (additional) Pfam domains to be included in the subsequent search, using a user-defined threshold. At the end of the run, the HMM profiles corresponding to the selected Pfam domains are downloaded and stored locally in a specific folder. If a list of PDB structures has been input, then the program will output files describing the content of these structures in terms of domains and their associated LBP's. The second script, *Genome_browsing.py* (bottom box), uses the output of the first script and requires as a further input the list of organisms to be analyzed. It creates a *Results* subfolder where all output files are moved to. The sequences of the proteins identified in each organism that contain at least one of the selected domains (and, if relevant, the associated LBP's) are stored in a subfolder with the organism's name (further separated in subfolders on a per domain basis). Summary files are also produced. (1): if ligands list has been submitted; (2): case-sensitive.

program.[42] The retrieved sequences are subjected to two filters: (i) for the presence of Pfam domains not included in the user's selection that match the same region of a selected domain with a better (i.e., lower) HMMER *E*-value; and (ii) for the presence of the LBP, if available. The computational flowchart is shown in Figure 1.

Retrieving_Domains.py. The main purpose of this script is to download the HMM's that describe the Pfam domains to be identified in the entire proteome sequences by the next script. These can be supplemented with LBP's, when relevant. This part of the procedure must thus start with assembling a list of the domains of interest. These can be: (i) directly input by the user, (ii) obtained from the analysis of sequences with known 3D structure, or (iii) both (Figure 1). Only in the case in which an user wants to extract the domains from the sequence of a protein of unknown structure, s/he should independently scan the sequence for Pfam domains using the interface at the Pfam Web site.[44] In (i) the user is asked to provide a list of Pfam domains, for which the script downloads the corresponding HMM's from the Pfam library. In (ii), the user inputs a list of PDB codes, whose protein sequences are downloaded from the PDB and scanned, using the HMMsearch function of HMMER 3.0, against the entire Pfam database to identify the domains they contain. If a PDB entry contains multiple chains, all the chains that are different in sequence are analyzed. Optionally, a list of ligands (identified by the three-letter chemical component identifier in the PDB database, corresponding to the HET field)[45] can be input, in which case the script will also identify the LBP and associate it to the Pfam domain within whose boundaries the amino acids of the LBP are (at least two amino acids must be within a domain to create an association; only the amino acids that are within the domain are then taken into account as the LBP). In (iii), the input data and results of (i) and (ii) are joined. Note that in (ii) and (iii), the user is asked to provide a threshold to decide whether the identification of a domain within a sequence is meaningful or not. This is done by setting an upper limit for the expectation value (*E*-value), which is a measure of the expected rate of errors in the identification of domains in protein sequences. Typical values are in the range 10^{-3}–10^{-5}. The results are optionally stored in separate subfolders (Figure 1). In all cases the *0Profiles* subfolder is created, which contains the HMM models downloaded as well as the log file of the script.

Genome_Browsing.py. The second main script, *Genome_browsing.py*, asks the user for a list of organisms of interest (Figure 1). The corresponding proteomes (which include all chromosomally encoded proteins as well as those encoded by plasmidic DNA) are downloaded from the NCBI ftp site[46] and then scanned for the occurrence of proteins containing the Pfam domains from the previous step. This is done by using the HMM models stored in the *0Profiles* subfolder with the HMMsearch function of HMMER 3.0. For the successful download of the proteomes, it is important that the name of the organisms is written exactly as listed at the NCBI, including, when relevant, the subspecies information (e.g., *Burkholderia xenovorans* LB400). A script is provided to obtain the lists of all prokaryotic and eukaryotic organisms whose full proteome is available, from which the names can be pasted (*Retrieving_organisms.py*).

Genome_browsing.py creates one folder per each organism in the list, in which all the sequences (in FASTA format) of the proteins that contain at least one of the domains of interest are saved. To reduce the rate of false positives, each sequence retrieved is compared against the whole Pfam database (Reverse-check in Figure 1). This allows the user to determine whether the domain of interest that has been identified in each retrieved sequence actually constitutes the best domain assignment for that region of the sequence. In other words, if a Pfam domain not in the list of the domains of interest matches better than any of the domains of the list a given region of the sequence (even though one of the domains of interest did match with an *E*-value better than the threshold), then the assignment of the protein as one of the pathway becomes dubious, and the sequence is therefore segregated for a possible further inspection. For domains that are associated to a LBP, the sequences are additionally filtered by requesting that they contain the LBP. A tolerance of 20% is applied to the spacing between amino acids in the LBP. Rejected files are moved in separate subfolders. A separate log file is created by this script, in addition to one which recapitulates all the results.

A tutorial for the use of the RDGB tool is given in the Supporting Information text and is included in the RDGB download file. At the end of the run, *Genome_browsing.py* will move all output files, including those generated by *Retrieving_Domains.py*, in the *Results* subfolder, which is created by the script (note that if *Results* already exists, then a directory *Results1* is created instead). The file and directory structure described above is maintained.

■ RESULTS AND DISCUSSION

The Strategy. The identification of biochemical pathways using computational methods has been the focus of a great deal of interest, especially since a large amount of sequence information for a variety of different organisms has been accumulating in genomic databases. Developments in the field have included identification of missing enzymes in otherwise complete pathways to lead experimental efforts for the discovery of new enzymes and gene functions and the annotation of the entire metabolic network of organisms, in a systems biology perspective. The latter is generally a quite complex task, which requires the application of sophisticated bioinformatics methods by highly skilled researchers. Another specific application is the comparison of the occurrence and distribution of a biochemical process in different organisms. This endeavor, which is computationally much less demanding than the aforementioned metabolic reconstructions, has a value, e.g., to determine how widespread a pathway for the acquisition of nutrients is or what pathways are shared by a group of pathogens (see for example our work on heme uptake as a source of iron for prokaryotes).[21] Although not computationally intensive, when performed on hundreds of organisms this kind of investigation generates a considerable amount of data, preventing manual inspection of all the results and therefore creating the need for a stable strategy that minimizes errors and is prone to automation.

In this work, we present a simple protocol to tackle the task mentioned above. The protocol relies on the identification of proteins on the basis of their domain content. This allows one to identify the possible homologues of all the proteins involved in a biochemical pathway of interest though a systematic scanning of the proteome (including the proteins that are encoded by both chromosomal and plasmidic DNA) of an organism. By inspecting the presence of at least one homologue for each protein in the pathway (or only of some selected key ones), it is possible to readily identify which organisms encode a pathway. Accordingly, one of the files output by RDGB consists of an Excel table reporting the occurrence of at least one homologue of each protein in the pathway (YES/NO) for all the organisms analyzed. At the same time, hints on variations on the composition of the pathway can also be obtained by analyzing more closely the organisms that lack only a small (with respect to the number of components in the pathway) number of proteins. Finally, the sequences identified can be inspected to ascertain possible differences in the mechanisms of substrate or intermolecular recognition, e.g., by identifying trends in amino acidic composition or the presence/absence of specific sequence motifs. Homology modeling can provide hints at the atomic level if the 3D structure of at least one representative of the family is available.

The present procedure thus starts with defining the list of domains that characterize the biochemical process of interest. We propose to use the Pfam library of domains because the annotation of the domains in the library is normally sufficiently detailed to allow users to evaluate the actual relevance of a domain to the biochemistry under investigation. The domains can practically be identified by scanning the sequence of one (or more) representative of each protein in the pathway against the full Pfam database with a reasonable *E*-value threshold (in the range $10^{-3}-10^{-5}$), using the service at the Pfam Web site.[44] When a protein binds a ligand/cofactor (e.g., organic ligands, metal ions, and metal-containing cofactors, such as heme) and the 3D structure of the bound form is available, it is possible to take advantage of the information on the protein−ligand mode of interaction to filter the results of domain-based searches. This information is condensed in the ligand binding pattern (LBP). The LBP defines the identity and the spacing of the amino acids in direct contact with or bound to the ligand; LBP's can be represented in the form AX*n*BX*m*C ..., where A, B, C, ... are the amino acids in contact with the ligand, and *n*, *m*, ... the number of amino acids in between two subsequent ligands. X is any amino acid.

After the list of domains (and associated LBP's) is compiled, it can be used to scan any complete proteome to identify the proteins that contain one (or more) of them. For domains associated with an LBP, the latter can be used to filter the results and improve the precision of the method. The filter is applied by imposing that the predicted protein contains all the ligands of the LBP with a spacing in sequence that it is maintained within ±20% (or ±1 amino acid for short spacing). This procedure leads to a significant reduction of the number of false positives (proteins wrongly predicted to be homologues of the ones of interest), as extensively documented for metal-binding proteins.[26] A further refinement to improve the precision is to

Table 1. Confusion Matrix and Performance Parameters for the RDGB Predictions against the Test Set for Heme Biosynthesis Described in Ref 47[a]

against the original data			
		condition	
		positive	negative
RDGB outcome	positive	312 (TP)	79 (FP)
	negative	8 (FN)	204 (TN)
after reclassifying the data			
		condition	
		positive	negative
RDGB outcome	positive	336 (TP)	55 (FP)
	negative	8 (FN)	204 (TN)
Performance Parameters			
sensitivity [TP/(TP + FN)]			97.7%
specificity [TN/(TN + FP)]			78.8%
precision [TP/(TP + FP)]			85.9%
accuracy [(TP + TN)/(TP + TN + FP + FN)]			89.6%

[a] The top part of the table measures the performance by assuming "expected but absent" genes in the reference data set as negatives; the central part of the table measures the performance after these data are reclassified as positives if a corresponding protein has been identified by RDGB. Performance parameters have been calculated only using the data after reclassification. TP: true positives; FP: false positives; TN: true negatives; and FN: false negatives. Sensitivity is the fraction of actual positives which are correctly identified as such; specificity is the fraction of actual negatives which are correctly identified as such; precision, also called positive predictive value, is the fraction of positive predictions which are correct; and accuracy is the fraction of predictions which are correct (both true positives and true negatives).

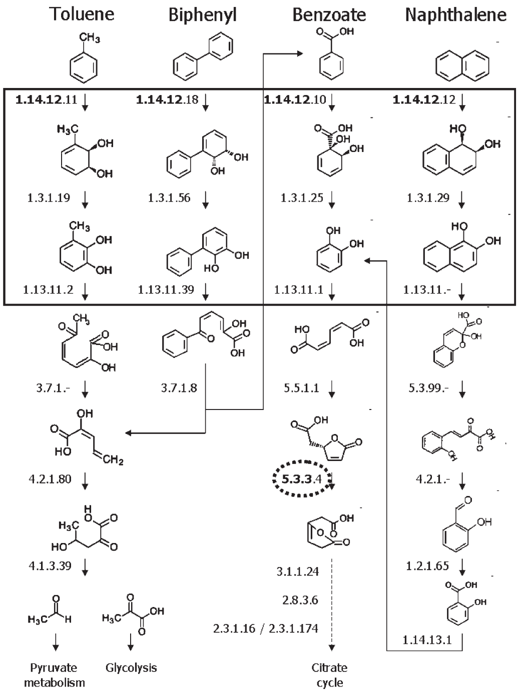


Figure 2. Overview of the pathways for the degradation of simple aromatic hydrocarbons, such as toluene, biphenyl, benzoate, and naphthalene. Note that naphthalene and biphenyl give rise to the formation of catechol and benzoate, which are further degraded in the benzoate pathway. Alternative pathways may exist, e.g., involving monooxygenation reactions. The regiochemistry of the ring-opening reaction generally depends on both the organism under consideration and the nature of the ring substituents. Subproducts, such as acetate, are not shown. This figure has been adapted from the KEGG pathway database. The benzoate degradation-specific enzyme muconolactone δ-isomerase is highlighted by a dotted ellipse.

check whether the region of the sequence corresponding to the domain of interest in each protein retrieved matches to another domain, not in the list, with a better (i.e., lower) *E*-value; if yes, then the protein is removed from the list of positives. This can happen because in the initial scan of the proteomes we only search for the domains of the list, in order to save time. By scanning the retrieved proteins, which typically are a very small fraction of each proteome, against the entire Pfam database, we can identify in each sequence other domains, not in the list, that overlap with the region spanned by the domain of interest (domains that correspond to protein regions not in overlap do not pose a problem and actually define multidomain proteins). Because both domains would match the sequence at an *E*-value lower than the threshold, it is useful not to discard the sequence immediately but rather to further inspect it. Useful guides are the extent of overlap between the two domains and the ratio of the corresponding *E*-values.

To validate the RDGB tool, we used it to recalculate automatically with default parameters a known test set. For the latter, we used the data described in a review on the heme biosynthesis pathway, where the occurrence of the relevant proteins in 67 prokaryotes has been investigated manually.[47] As already mentioned, for each protein in the pathway, we define as "positive" the case in which at least one homologue is present in an organism's proteome and as "negative" otherwise (Supporting Information, Table S1). The total number of data thus equals the product of the number of proteins in the pathway and the number of organisms studied (9 × 67 = 603 in the present test set). The sensitivity of our method is 97.7%, whereas its specificity is 78.8%, and the overall accuracy is 89.6% (Table 1). Note that the specificity of RDGB is not outstanding. This is caused by the number of false positives retrieved. As described also for the examples in the next sections, this is what should be expected as a consequence of the inclusion of domains with a catalytic function that is not specific of the pathway of interest, such as, for the present test case, HemG and HemY (whose Pfam domains, respectively, describe flavodoxins and flavin-containing amine oxidoreductases). The latter indeed account for ca. 80% of the false positives. Thus, domains that are not specific should preferably not be used to decide whether an organism is endowed with a given biochemical pathway. Rather, they should be used to confirm the results of a RDGB analysis carried out using specific domains. On the other hand, it is noteworthy that the application of RDGB leads to the identification of proteins that were expected but not detected in ref 47. For example, the HemD protein, which should be present in all tetrapyrrole-synthesizing prokaryotes, had been described as missing in 11 organisms, for 10 of which RDGB was instead able to identify it.

An Example Application to the Degradation of Aromatic Hydrocarbons. To demonstrate an application of the methodology described here (Figure 1), we characterized the pathways for the aerobic degradation of aromatic hydrocarbons that start with cis-dihydroxylation of the substrate in 1136 prokaryotic proteomes available from the NCBI database. As a further difficulty of analysis, the enzymes in these pathways can be encoded both by chromosomal and plasmidic genes. This does not pose a problem with RDGB, as our tool analyses all the proteins of the organism regardless of their genetic origin. The substrates of interest in the present group of processes range from toluene to biphenyl and naphthalene, including various other compounds in which the aromatic ring(s) are differently substituted. Figure 2 presents a general overview of these pathways, as it can be derived from the information in the KEGG database.[17] Note that many variants to these pathways can exist in nature, e.g., regarding the regiochemistry of some reactions or the involvement of mono-oxygenation reactions.[32] It can be seen that only the upper part of the pathways is common to all substrates, i.e., the initial dihydroxylation, followed by dehydrogenation and then by the opening of one aromatic ring through the cleavage of a carbon—carbon double bond (Figure 2). Here we are dealing with three of the four known families of hydroxylating dioxygenases, namely with the so-called toluene/biphenyl, naphthalene, and benzoate dioxygenases[34] (the substrate specificity of these enzymes is much broader than the names suggest), which are heteromultimers consisting of α- and β-subunits. A fourth class of dioxygenases exists, namely phthalate dioxygenases (which include also carbazole and 2-oxo-1,2-dihydroquinoline among their substrates), which are instead homomultimers with an α_n quaternary structure. Phthalate dioxygenases differ significantly in sequence from the members of the other three families[34,48] and are actually associated to a different Pfam domain. Among the enzymes (and their ancillary proteins, such as electron-transporting ferredoxins) of Figure 2, there are only two domains that are specific to the pathways of interest, namely the ring_hydroxyl_A and ring_hydroxyl_B domains. These are contained respectively in the α- and β-subunits of the initial dihydroxylating dioxygenase (Table 2). The other proteins instead contain relatively common functional domains that are present also in enzymes involved in other metabolic processes. Further along the degradation pathways of Figure 2, another pathway-specific domain is in muconolactone δ-isomerase, which plays a role within the degradation of catechol. The latter can be generated during the degradation of either benzoate or naphthalene or can be present itself in the environment. With these three specific domains (Table 2), we retrieved a total of 1099 proteins (Table 3).

The proteome of an organism able to aerobically degrade aromatic hydrocarbons must encode at least one homologue for all the proteins in the toluene pathway of Figure 2, which is the simplest of those analyzed here. In particular, because all proteins but those contained in the first enzyme of the pathway are common to various metabolic processes, it is the presence of both ring_hydroxyl_A- and ring_hydroxyl_B-containing proteins that can be used as the indicator of the ability to aerobically degrade toluene. With this simple rule, 919 organisms were found to be unable to aerobically degrade aromatic hydrocarbons, whereas 178 organisms were found to possess the right enzymatic portfolio. In 14 out of 53 cases where either a ring_hydroxyl_A- or a ring_hydroxyl_B-containing protein was missing, we found out that a suitable protein was detected with an *E*-value just above the chosen threshold, so these organisms were included in the list of putative degraders (shown as yellow cells in Supporting Information, Table S2). Instead, two organisms missed one or more unspecific Pfam domains, so we assigned them as unable to perform the process. The organisms where we could identify only one out of the two subunits of the initial ring-hydroxylating dioxygenase (Figure 2) were marked with "?" in Supporting Information, Table S2.

The detection of all other domains involved in the pathway, regardless of their specificity, can be taken as a useful countercheck that there are no major problems with the selection of the pathway-specific domains. To this end, we checked as an example the toluene pathway of Figure 2, including the additional relevant, unspecific domains (Table 3). Overall, we thus retrieved a total of 65 196 proteins, corresponding to 65 762 hits to the selected Pfam domains (Table 2). All the hits are reported in the Supporting Information (Supplementary Table S3). In all cases but one, the proteome of an organism encoding ring_hydroxyl_A- and

Table 3. Number of Hits and Number of Organisms Where Each Pfam Domain of Table 2 Has Been Identified

Pfam domain	protein	no. of hits	no. of organisms
Ring_hydroxyl_A	Dioxygenase	482	200
Ring_hydroxyl_B		461	185
Rieske		3217	726
Adh_short	Dehydrogenase	32 539	1108
Glyoxalase	Dioxygenase	5015	887
Abhydrolase_1	Hydrolase	17 847	1098
FAA_hydrolase	Dehydratase	2854	788
HMGL-like	Aldolase	3191	988
MIase	Isomerase	156	140

Table 2. Proteins Involved in the Aerobic Biodegradation of Aromatic Hydrocarbons (Figure 2)[a]

protein	EC number	Pfam	PDB	ligand
Dioxygenase	1.14.12	**Ring_hydroxyl_A**, Rieske	1ULI; 2B1X; 2BMO; 2GBW; 2HMJ; 3EN1	Fe, FeS
	1.14.12	**Ring_hydroxyl_B**	–	–
Isomerase	5.3.3	**MIase**	–	–
Dehydrogenase	1.3.1	Adh_short	–	–
Dioxygenase	1.13.11	Glyoxalase	1EIQ; 1HAN; 1KMY; 2EHZ; 2EI2; 2ZI8; 3HPV	Fe
Hydrolase	3.7.1	Abhydrolase_1	–	–
Dehydratase	4.2.1	FAA_hydrolase	–	–
Aldolase	4.1.3	HMGL_like	–	–

[a] The proteins in the top part of the table contain domains specific to the pathways of interest (in bold), whereas the proteins in the bottom part are not specific. For each protein, we report the EC number (only three levels are given, as the fourth depends on the identity of the substrate), the composition in Pfam domains, and the corresponding PDB structures (only if a ligand is present and thus an LBP is available).

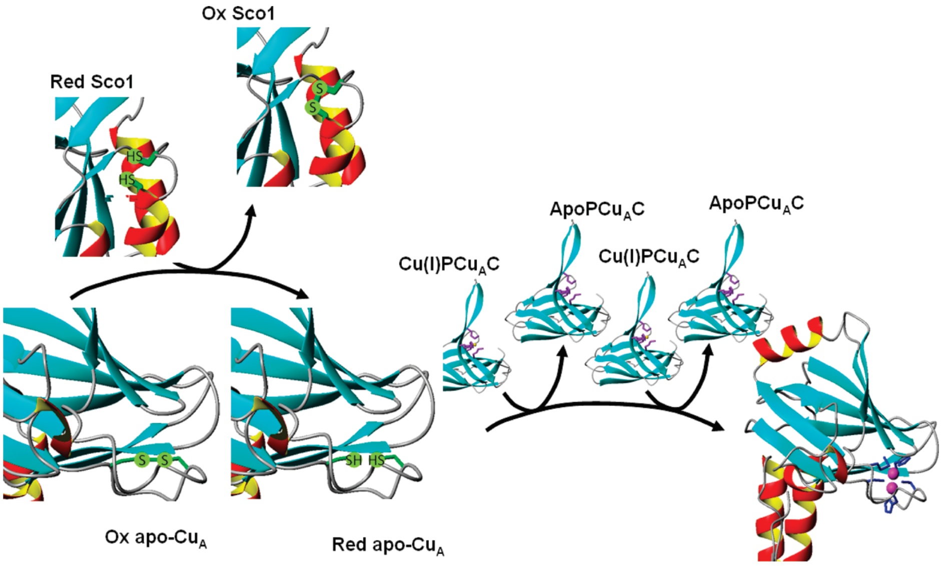


Figure 3. The mechanism of assembly of the CuA cofactor, as described in ref 37 (Ox = oxidized and Red = reduced). The sulfur atoms of redox-active cysteines are shown as circles. Metal ligands are shown as sticks. Metal ions are shown as spheres.

Table 4. Number of Hits and Number of Organisms Where Domains Related to the Mechanism of Assembly of the Cu$_A$ Cofactor (Figure 3) Were Detected

Pfam domain	protein	no. of hits	no. of organisms
COX2	Cox2	899	548
SCO1-SenC	Sco1	790	456
DUF461	PCu$_A$C	454	366

ring_hydroxyl_B-containing proteins contained also all other domains (Supporting Information, Table S2).

Finally, the detection of the MIase domain, which uniquely identifies muconolactone δ-isomerase (EC 5.3.3.4) in the benzoate pathway (Figure 2), permits the identification of organisms that can potentially degrade also biphenyls, benzoate, and naphthalene (all of them or a combination thereof). This domain has been detected in 140 organisms. Of these, 23 corresponded to organisms that were assigned as unable to degrade hydrocarbons and 5 to organisms marked with "?" in Supporting Information, Table S2. The corresponding genes are in the neighborhood of proteins annotated as hypothetical, suggesting the presence of uncharacterised mechanisms for the degradation of the substrates of interest here or, alternatively, that the MIase domain can play other uncharacterised roles.

An Example Application to Assembly of the CuA Cofactor. Cytochrome *c* oxidases use the Cu$_A$ cofactor as the entry point of the electron that is delivered by cytochrome *c* into the enzyme. Cu$_A$ is a dinuclear copper site contained in subunit II of the enzyme (Cox2), whose correct assembly is crucial for enzyme function. It has been shown by NMR that in *Thermus thermophilus* the assembly process is mediated by the soluble metallochaperone PCu$_A$C and the Sco1 thiol-disulfide reductase, which maintains the Cys residues in the Cu$_A$ binding site of Cox2 in the reduced state[37] (Figure 3). We used this relatively small ensemble of proteins for a further demonstration of an application of RDGB. Also nitrous oxide reductases (NosZ) contain a Cu$_A$-binding domain that is homologous to that of Cox2; the assembly of the Cu$_A$ cofactor in NosZ has not been studied in detail, but it is likely to involve a mechanism similar to Cox2. The data are reported in Supporting Information, Table S4.

Among the same prokaryotic organisms analyzed in the previous section 548 contained enzymes with a soluble Cu$_A$-binding domain (Table 4), corresponding to 48.3% of the whole ensemble. The occurrence of PCu$_A$C and Sco1 homologues was less frequent, corresponding respectively to 32.3 and 40.2% of the organisms analyzed. It is relevant to address the co-occurrence of these proteins; 283 organisms contained all three proteins, corresponding to 24.9% of the data set. In other words, one-quarter of the prokaryotic organisms investigated encoded in their proteomes a Cu$_A$-binding domain, most likely in Cox2 as this is much more widespread than NosZ,[24] and the two accessory proteins that have been demonstrated to be active in the assembly of the cofactor in *T. thermophilus*. This corresponds to 51.5% of the organisms that encode at least one Cu$_A$-binding domain. In 57 cases (5.0% of the organisms), a Cu$_A$-containing domain could not be detected but Sco1 (1.0%) or PCu$_A$C (0.2%) or both (3.8%) were contained in the proteome. The occurrence of Sco1 in the absence of any Cu$_A$-containing enzyme had been noted before and was proposed to be due to its possible activity as a thiol-disulfide oxidoreductase.[24] However, the relatively common occurrence of both Sco1 and PCu$_A$C in the absence of any Cu$_A$-containing enzyme may suggest that the assembly mechanism of the Cu$_A$ cofactor, or a close variant of it, may be relevant also for the assembly of other cuproenzymes. Finally, it is worth noting that 108 organisms (9.5%) encode a Cu$_A$-containing enzyme while lacking both Sco1 and PCu$_A$C. Thus, some yet uncharacterised assembly mechanisms may be operative in organisms such as Mycobacteria (and various other Actinobacteria), δ-proteobacteria, and Cyanobacteria.

■ CONCLUSIONS

Following the strategy of Figure 1, the RDGB computational tool described here allows users to characterize known metabolic pathways in a wide range of organisms, exploiting three main databases: Pfam, PDB, and NCBI. The functional (Pfam) domains are taken as characterizing elements of the proteins of interest and can be obtained also from the analysis of 3D structures available from the PDB. In addition, by investigating the PDB for the presence of ligands, it is possible to obtain one or more ligand binding patterns (LBP's) that are associated to the Pfam functional domain. The LBP can be used as a filter to reduce the number of false positives. The sequences of all the proteins encoded by both the chromosomal and plasmidic DNA of an organism with fully sequenced genome are obtained from the NCBI. The data output by RDGB can also be used to identify possible variants of known pathways, such as the occurrence of alternative steps, or, at the atomic level, of different modes of intermolecular interaction with the substrate.

The performance of RDGB has been validated against a known test set. Furthermore, the overall approach has been demonstrated through two examples. In one, we analyzed the processes for aerobic degradation of mono- and poly-cyclic aromatic hydrocarbons, which include toluene, naphthalene, and biphenyls. Out of 1136 organisms analyzed, only 178 were able to degrade at least one of the above compounds; 112 could potentially degrade all of them; and 39 organisms missed only one requested protein/subunit along the pathway, preventing us from assigning them as degraders or nondegraders. In the second example, we investigated the distribution of some accessory proteins that are involved in the assembly of the Cu_A cofactor. The data indicate that one or more Cu_A-binding domains can be detected in nearly half of the organisms analyzed. Among the organisms encoding Cu_A-binding domains, 51.5% contain both Sco1 and PCu_AC, demonstrating that this mechanism of Cu_A biogenesis is quite widespread in prokaryotes.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information. Tables summarizing the RDGB results for the test set, results obtained by investigating the per organism distribution of the proteins involved in the biodegradation of aromatic hydrocarbons, output log file produced by RDGB for all the proteins involved in the biodegradation of aromatic hydrocarbons in all the investigated organisms, and results obtained by investigating the per organism distribution of the proteins involved in the assembly of the Cu_A cofactor. Instructions for the installation and the use of RDGB. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: ivanobertini@cerm.unifi.it.

■ ACKNOWLEDGMENT

Financial support from MIUR (FIRB-FUTURO IN RICERCA 2008, RBFR08WGXT and FIRB-Proteomica RBRN07BMCT) is gratefully acknowledged.

■ REFERENCES

(1) Enright, A. J.; Iliopoulos, I.; Kyrpides, N. C.; Ouzounis, C. A. Protein interaction maps for complete genomes based on gene fusion events. *Nature* **1999**, *402*, 86–90.

(2) Marcotte, E. M.; Pellegrini, M.; Ng, H. L.; Rice, D. W.; Yeates, T. O.; Eisenberg, D. Detecting protein function and protein-protein interactions from genome sequences. *Science* **1999**, *285*, 751–753.

(3) Overbeek, R.; Fonstein, M.; D'Souza, M.; Pusch, G. D.; Maltsev, N. The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 2896–2901.

(4) Wolf, Y. I.; Rogozin, I. B.; Kondrashov, A. S.; Koonin, E. V. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res.* **2001**, *11*, 356–372.

(5) Snel, B.; Bork, P.; Huynen, M. A. The identification of functional modules from the genomic association of genes. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 5890–5895.

(6) Pellegrini, M.; Marcotte, E. M.; Thompson, M. J.; Eisenberg, D.; Yeates, T. O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 4285–4288.

(7) Pagel, P.; Wong, P.; Frishman, D. A domain interaction map based on phylogenetic profiling. *J. Mol. Biol.* **2004**, *344*, 1331–1346.

(8) Snel, B.; Lehmann, G.; Bork, P.; Huynen, M. A. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.* **2000**, *28*, 3442–3444.

(9) von Mering, C.; Huynen, M.; Jaeggi, D.; Schmidt, S.; Bork, P.; Snel, B. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* **2003**, *31*, 258–261.

(10) Lee, I.; Date, S. V.; Adai, A. T.; Marcotte, E. M. A probabilistic functional network of yeast genes. *Science* **2004**, *306*, 1555–1558.

(11) Date, S. V.; Marcotte, E. M. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat. Biotechnol.* **2003**, *21*, 1055–1062.

(12) Osterman, A.; Overbeek, R. Missing genes in metabolic pathways: a comparative genomics approach. *Curr. Opin. Chem. Biol.* **2003**, *7*, 238–251.

(13) Cordwell, S. J. Microbial genomes and "missing" enzymes: redefining biochemical pathways. *Arch. Microbiol.* **1999**, *172*, 269–279.

(14) Gianchandani, E. P.; Brautigan, D. L.; Papin, J. A. Systems analyses characterize integrated functions of biochemical networks. *Trends Biochem. Sci.* **2006**, *31*, 284–291.

(15) Price, N. D.; Reed, J. L.; Palsson, B. O. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* **2004**, *2*, 886–897.

(16) Tyson, J. J.; Chen, K.; Novak, B. Network dynamics and cell physiology. *Nat. Rev. Mol. Cell Biol.* **2001**, *2*, 908–916.

(17) Kanehisa, M.; Araki, M.; Goto, S.; Hattori, M.; Hirakawa, M.; Itoh, M.; Katayama, T.; Kawashima, S.; Okuda, S.; Tokimatsu, T.; Yamanishi, Y. KEGG for linking genomes to life and the environment. *Nucleic Acids Res.* **2008**, *36*, D480–D484.

(18) Kanehisa, M.; Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30.

(19) Karp, P. D.; Ouzounis, C. A.; Moore-Kochlacs, C.; Goldovsky, L.; Kaipa, P.; Ahren, D.; Tsoka, S.; Darzentas, N.; Kunin, V.; Lopez-Bigas, N. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.* **2005**, *33*, 6083–6089.

(20) Matthews, L.; Gopinath, G.; Gillespie, M.; Caudy, M.; Croft, D.; de Bono, B.; Garapati, P.; Hemish, J.; Hermjakob, H.; Jassal, B.; Kanapin, A.; Lewis, S.; Mahajan, S.; May, B.; Schmidt, E.; Vastrik, I.; Wu, G.; Birney, E.; Stein, L.; D'Eustachio, P. Reactome knowledgebase of human biological pathways and processes. *Nucleic Acids Res.* **2009**, *37*, D619–D622.

(21) Cavallaro, G.; Decaria, L.; Rosato, A. Genome-based analysis of heme biosynthesis and uptake in prokaryotic systems. *J. Proteome. Res.* **2008**, *7*, 4946–4954.

(22) Bertini, I.; Cavallaro, G.; Rosato, A. Evolution of mitochondrial-type cytochrome *c* domains and of the protein machinery for their assembly. *J. Inorg. Biochem.* **2007**, *101*, 1798–1811.

(23) Sharma, S.; Cavallaro, G.; Rosato, A. A systematic investigation of multi-heme c-type cytochromes in prokaryotes. *J. Biol. Inorg. Chem.* **2010**, *15*, 559–571.

(24) Banci, L.; Bertini, I.; Cavallaro, G.; Rosato, A. The functions of Sco proteins from genome-based analysis. *J. Proteome Res.* **2007**, *6*, 1568–1579.

(25) Bertini, I.; Cavallaro, G.; Rosato, A. Cytochrome c: occurrence and functions. *Chem. Rev.* **2006**, *106*, 90–115.

(26) Andreini, C.; Bertini, I.; Rosato, A. Metalloproteomes: a bioinformatic approach. *Acc. Chem. Res.* **2009**, *42*, 1471–1479.

(27) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410.

(28) Hulsen, T.; Huynen, M. A.; de Vlieg, J.; Groenen, P. M. Benchmarking ortholog identification methods using functional genomics data. *Genome Biol.* **2006**, *7*, R31.

(29) Claudel-Renard, C.; Chevalet, C.; Faraut, T.; Kahn, D. Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res.* **2003**, *31*, 6633–6639.

(30) *Priority pollutants*; United States Environmental Protection Agency: Washington, D.C.; http://water.epa.gov/scitech/swguidance/methods/pollutants.cfm. Accessed January 17, 2011.

(31) Ramos, J. L.; Diaz, E.; Dowling, D.; de Lorenzo, V.; Molin, S.; O'Gara, F.; Ramos, C.; Timmis, K. N. The behavior of bacteria designed for biodegradation. *Biotechnology (N.Y.)* **1994**, *12*, 1349–1356.

(32) Cao, B.; Nagarajan, K.; Loh, K. C. Biodegradation of aromatic compounds: current status and opportunities for biomolecular approaches. *Appl. Microbiol. Biotechnol.* **2009**, *85*, 207–228.

(33) Yakimov, M. M.; Timmis, K. N.; Golyshin, P. N. Obligate oil-degrading marine bacteria. *Curr. Opin. Biotechnol.* **2007**, *18*, 257–266.

(34) Gibson, D. T.; Parales, R. E. Aromatic hydrocarbon dioxygenases in environmental biotechnology. *Curr. Opin. Biotechnol.* **2000**, *11*, 236–243.

(35) Vaillancourt, F. H.; Bolin, J. T.; Eltis, L. D. The ins and outs of ring-cleaving dioxygenases. *Crit. Rev. Biochem. Mol. Biol.* **2006**, *41*, 241–267.

(36) Carr, H. S.; Winge, D. R. Assembly of Cytochrome c Oxidase within the Mitochondrion. *Acc. Chem. Res.* **2003**, *36*, 309–316.

(37) Abriata, L. A.; Banci, L.; Bertini, I.; Ciofi-Baffoni, S.; Gkazonis, P.; Spyroulias, G. A.; Vila, A. J.; Wang, S. Mechanism of Cu(A) assembly. *Nat. Chem. Biol.* **2008**, *4*, 599–601.

(38) *Python Documentation*, v2.7.1; Global Module Index; Python Software Foundation: Wolfeboro Falls, NH; http://docs.python.org/modindex.html. Accessed January 17, 2011.

(39) Bateman, A.; Coin, L.; Durbin, R.; Finn, R. D.; Hollich, V.; Griffiths-Jones, S.; Khanna, A.; Marshall, M.; Moxon, S.; Sonnhammer, E. L.; Studholme, D. J.; Yeats, C.; Eddy, S. R. The Pfam protein families database. *Nucleic Acids Res.* **2004**, *32*, database issue, D138−D141.

(40) Sonnhammer, E. L.; Eddy, S. R.; Durbin, R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **1997**, *28*, 405–420.

(41) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(42) Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **1998**, *14*, 755–763.

(43) Pruitt, K. D.; Tatusova, T.; Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **2005**, *33*, D501–D504.

(44) *Pfam sequence search*; The Wellcome Trust, Sanger Institute: London; http://pfam.sanger.ac.uk/search. Accessed January 17, 2011.

(45) *RCSB PDB; PDB Chemical Component Dictionary Format Description*; Rutgers and University of California, San Diego: New Brunswick, NJ and San Diego, CA; http://deposit.pdb.org/cc_dict_tut.html. Accessed January 17, 2011.

(46) *Genomes ftp site* National Center for Biotechnology Information: Bethesda, MD; ftp.ncbi.nih.gov/genomes/. Accessed January 17, 2011.

(47) Panek, H.; O'Brian, M. R. A whole genome view of prokaryotic haem biosynthesis. *Microbiology* **2002**, *148*, 2273–2282.

(48) Nam, J. W.; Nojiri, H.; Yoshida, T.; Habe, H.; Yamane, H.; Omori, T. New classification system for oxygenase components involved in ring-hydroxylating oxygenations. *Biosci. Biotechnol. Biochem.* **2001**, *65*, 254–263.