D468–D472 *Nucleic Acids Research, 2010, Vol. 38, Database issue* *Published online 11 November 2009*
doi:10.1093/nar/gkp1001

MetaBioME: a database to explore commercially useful enzymes in metagenomic datasets

Vineet K. Sharma, Naveen Kumar, Tulika Prakash and Todd D. Taylor*

MetaSystems Research Team, Computational Systems Biology Research Group, Advanced Computational Sciences Department, Advanced Science Institute, RIKEN, Yokohama, Kanagawa 230-0045, Japan

Received August 15, 2009; Revised October 8, 2009; Accepted October 17, 2009

ABSTRACT

Microbial enzymes have many known applications as biocatalysts in biotechnology, agriculture, medical and other industries. However, only a few enzymes are currently employed for such commercial applications. In this scenario, the current onslaught of metagenomic data provides a new unexplored treasure trove of genomic wealth that can not only enhance the enzyme repertoire by the discovery of novel commercially useful enzymes (CUEs) but can also reveal better functional variants for existing CUEs. We prepared a catalogue of CUEs using text mining of PubMed abstracts and other publicly available information, and manually curated the data to identify 510 CUEs. Further, in order to identify novel homologues of these CUEs, we identified potential ORFs in publicly available metagenomic datasets from 10 diverse sources. Using this strategy, we have developed a resource called MetaBioME (http://metasystems.riken.jp/metabiome/) that comprises (i) a database of CUEs and (ii) a comprehensive platform to facilitate homology-based computational identification of novel homologous CUEs from metagenomic and bacterial genomic datasets. Using MetaBioME, we have identified several novel homologues to known CUEs that can potentially serve as leads for further experimental verification.

INTRODUCTION

Characteristics such as high efficiency and stereoselectivity render naturally occurring enzymes suitable for commercial applications (1). These 'commercially useful enzymes' (or CUEs), predominantly used as 'biocatalysts', offer ecologically friendly or 'green' solutions for the implementation of biochemical processes at a reduced cost and produce a large variety of chemical substances (2). Despite these merits, only a limited number of enzymes have been commercially exploited. This limitation is primarily due to the lack of availability of microbial enzymes that can perform the desired chemical reactions.

In the current era of sequencing, when numerous genomes are being sequenced, the discovery of novel ORFs far exceeds the rate of their functional characterization, resulting in most of these ORFs being labelled as hypothetical proteins with unknown functions. The wetlab work to be invested into enzyme characterization is a much more tedious process and is one of the reasons for the growing gap between the number of potential enzymes deduced from these genomic sequences and those actually implemented in industry.

Another reason for the above limitation is that the CUEs that are currently used may not be 'ideal' enzymes for a given bioprocess, and sometimes the industrial processes have to be designed to purposely fit these mediocre enzymes (3). Therefore, improving the existing enzymes to make them suitable for commercial exploitation or finding better functional variants is a key challenge.

One promising approach is to augment our knowledgebase by exploring the inherent diversity of nature that harbours numerous species and their constituent enzymes that perform numerous transformations of molecules in diverse biological systems with great precision and specificity. In this scenario, metagenomics has emerged as a powerful culture-independent approach for exploring the complexity and diversity of microbial genomes in their natural environments (4). Potentially, it can not only enhance the enzyme repertoire by the discovery of novel CUEs but can also reveal better functional variants for the existing CUEs. The current onslaught of metagenomic data provides a unique opportunity to discover novel functional variants for existing CUEs using sequence homology-based approaches.

Therefore, in the present work, to first catalogue the known CUEs, we used publicly available information to curate a unique and comprehensive database of CUEs mostly comprising biocatalysts currently used in diverse commercial applications or having potential applications.

*To whom correspondence should be addressed. Tel: +81-45-503-9285; Fax: +81-45-503-9176; Email: taylor@riken.jp

Further, in order to find the homologues of these CUEs, we explored 10 metagenomic data sources and 971 completed bacterial genomes and identified several novel homologues for most of the known CUEs. Using this strategy, we developed the comprehensive Metagenomic BioMining Engine (MetaBioME), which can be used as an intuitive search engine to access manually curated data on the CUEs, stored in a relational database, along with several options to identify their homologues from multiple metagenomic datasets and completed bacterial genomes.

DATABASE CONSTRUCTION AND CONTENTS

Enzyme database

For this analysis, we have exclusively used the Enzyme Commission number (EC number) system to refer to enzymes and define their functions (5). Information on the complete set of 4877 enzymes annotated with EC numbers was retrieved from the ENZYME nomenclature database, as available at ExPASy (March 3, 2009) (6). The corresponding Swiss-Prot sequences were retrieved from the Swiss-Prot database (release 56.9, March 3, 2009) (7).

Database of CUEs

We curated a database of 510 enzymes with known or potential commercial applications (CUEs) using the information available at NCBI PubMed (8) and BRENDA (9). All 'English' abstracts containing the keyword 'enzyme' were retrieved from PubMed in XML format and imported into a MySQL database (version 5.1) (Figure 1). The initial set of candidate CUEs were identified using the 'Natural Language full-text search' and 'Boolean full-text search' features of MySQL. Additional information on known CUEs was retrieved from BRENDA. Taken together, these candidate CUEs were manually curated to identify the final set of 510 CUEs (CUEsDB). Based on their known application, these CUEs were classified into nine broad application categories, namely: Agriculture, Biosensor, Biotechnology, Energy, Environment, Food and Nutrition, Medical, Other Industries and Miscellaneous.

Other resources

The non-redundant (NR) sequence database, sequences of 971 completed bacterial genomes (ftp.ncbi.nih.gov/genomes/Bacteria as of September 21, 2009), and Conserved Domain Database (ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd) were retrieved from NCBI (8,10). The Protein Data Bank (PDB) database was retrieved from the Worldwide Protein Data Bank (wwPDB) (http://www.wwpdb.org/) (11). Protein structures were created using Rasmol (version 2.6) (12).

Mining the metagenomic databases

In the current version of the database, we have included the publicly available metagenomic sequence data from 10 sources (environments) comprising 44 datasets (details are available at http://metasystems.riken.jp/metabiome/metagenome.php) generated using Sanger sequencing technology except in the case of mouse gut where both Sanger and 454 sequencing technologies were used. The assembled metagenomic data was retrieved from NCBI Entrez Genome Project (http://www.ncbi.nlm.nih.gov/genomes/lenvs.cgi). Complete and partial ORFs ($\geq$150 nucleotides) were predicted in the metagenomic contigs using the SuperGene algorithm (part of our in-house iMetaSys pipeline (13,14)) that integrates both the Glimmer (15) and MetaGene (16) gene prediction software. The Cd-hit program (version 3.1.2) (17) was used to cluster the metagenomic ORFs. Swiss-Prot protein sequences were available for only 409 CUEs and

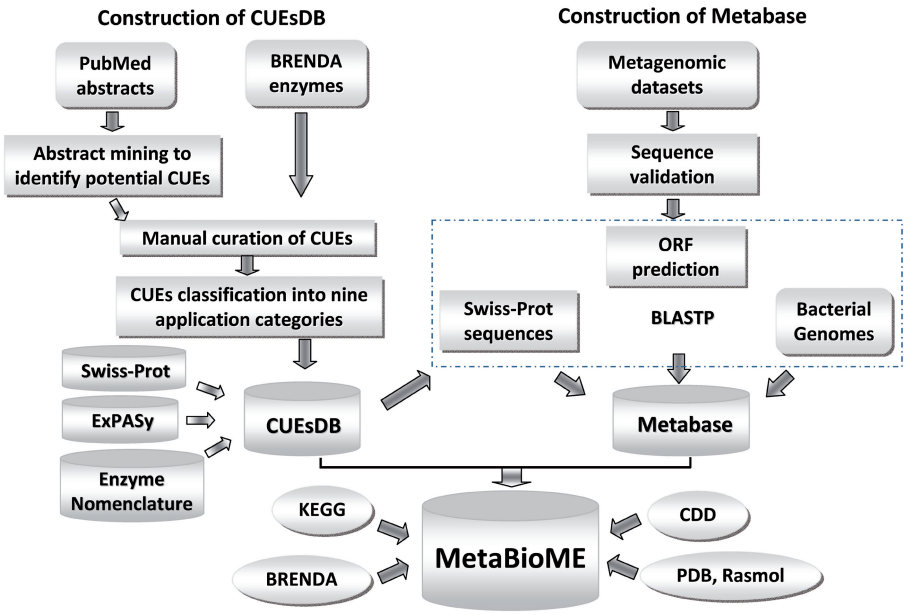

Figure 1. Steps in the construction of the MetaBioME database.

these sequences were aligned with the predicted metagenomic ORFs for each metagenomic dataset using BLASTP with a threshold of $E < 10^{-6}$. The output was generated in XML format, parsed and imported into a MySQL database (Metabase).

Web Interface and Metabase development

Open Source LAMP (Red Hat Enterprise Linux 4) Technology, Apache (version 2.2.8), MySQL (version 5.0.45), PHP (version 5.2.4) and Perl (version 5.8.5)) were used for development of the GUI and back-end database called 'Metabase'. The web-server was developed using the Apache HTTP Server (version 2.2.8). Client-side scripting was done using XHTML, JavaScript and AJAX, and server-side scripting was done using PHP and XML. The external applications BLAT (v34) (18), BLAST (version 2.2.17) and MAFFT (version 6.240) (19) were integrated for analysis.

RESULTS, QUERIES AND WEB INTERFACE

Distribution of CUEs in application categories

The distribution of 510 CUEs in nine application categories provides a useful schema for selecting enzymes involved in an application area of interest. Since an enzyme may be employed in more than one application, some overlaps exist in the distribution of CUEs in these nine categories. Among these categories, the highest number of CUEs are present in the 'Biotechnology' category (234, 46%) and the lowest number are in the 'Energy' category (13, 3%) (Supplementary Figures S1–S3).

Identification of potential homologues to known CUEs

Using a homology-based approach, we identified 199 (49%) novel homologues for known CUEs in the metagenomic datasets using a stringent threshold of identity $\geq$50% and coverage $\geq$90% (Table 1). Upon relaxing the above cut-off (identity $\geq$30% and coverage $\geq$90%), we identified an expanded list of novel homologues for a total of 305 (75%) out of 409 CUEs in the metagenomic datasets (Supplementary Table S1). Within this expanded list, homologues for 20 CUEs were commonly found in all nine metagenomic datasets (the coral viral metagenome dataset was excluded from this analysis), while homologues for 64 CUEs only appeared once each among the nine metagenomic datasets (Supplementary Table S2).

Description of web resource: MetaBioME

MetaBioME has two main components (i) a curated database of CUEs and (ii) comprehensive bio-mining options to search for novel homologues to the known CUEs in metagenomic and bacterial genomic datasets. For comprehensive querying, we have designed four query pages that are briefly described below.

MetaSearch: search for homologous CUEs in metagenomic datasets and completed bacterial genomes

The 'MetaSearch' query page is designed to identify novel homologues to the existing set of CUEs from multiple metagenomic datasets and completed bacterial genomes. It consists of a set of CUEs pre-classified in nine application categories that help the user to select any CUEs of interest based on the area of application (Supplementary Figure S4). Queries can be made by selecting one or more of the application categories or by using the 'Advanced Search' option to select any particular enzyme class (EC) or enzyme name (Supplementary Figure S5). This selected set of enzymes can be searched for in the available metagenomic data or completed bacterial genomes. Queries can also be made using multiple keywords and Boolean operators by selecting different attributes, such as enzyme name or keywords, biochemical pathway and substrates or products. A sample query to search for CUEs belonging to the 'Environment' application category in the 'Soil' metagenomic source is shown in Supplementary Figures S4, S6 and S7.

On query submission, MetaBioME examines the sequence similarity of all known Swiss-Prot sequences of CUEs belonging to the selected application categories with all the predicted metagenomic ORFs of the selected metagenomic dataset(s) or with all proteins of the selected bacterial genomes. The subsequent 'MetaResults' page displays the qualified hits as a table sorted on the basis of percent coverage (Supplementary Figure S6). Comprehensive information can be retrieved by clicking on the Swiss-Prot ID link on the MetaResults page, opening up the 'MetaBioME profile' page (Supplementary Figure S7). The profile page summarizes various information about the selected CUE. This is followed by a table of all the predicted ORFs in the metagenomic contig or bacterial genome with the description of the ORF that showed the highest similarity to the selected Swiss-Prot sequence of the CUE, also displayed in the contig view window. This is followed by an alignment view of the homologous ORF with the CUE sequence, a summary of the closest match of the homologous ORF to a known finished bacterial genome, and information on the closest available PDB structure. The alignment of the CUE sequence with all novel metagenomic ORFs (all datasets) clustered using cd-hit is displayed in the next window. A list of other Swiss-Prot IDs belonging to the same EC number that showed lower similarity is shown in the next table.

'MetaBioME Rating' rates the homologous ORF on a scale of 1–5 stars (weakest to best match). In the case of a good match ($\geq$2 stars), users can perform an 'Advanced Analysis' such as (i) examine the alignment of the CUE sequence with the homologous ORF, (ii) examine the sequence similarity among all Swiss-Prot sequences of the CUE and the homologous ORF, (iii) examine the presence of conserved domains in the homologous ORF using the NCBI Conserved Domain Database (CDD) or (iv) look for more homologues of the CUE in other metagenomic datasets or bacterial genomes.

Table 1. Distribution of CUEs showing significant homology with novel ORFs in metagenomic datasets

Application	[a]Total potential homologs (%)	Homologous ORFs predicted in Metagenomic Datasets (Coverage: ≥90%, Identity ≥50%)[b]								
		Human gut	Mouse gut	Termite gut	Marine	Mine drainage	Sludge	Soil	Microbial mat	Whale fall
Agriculture	14 (34)	11	1	4	12	3	7	5	4	6
Biosensor	34 (51)	27	1	3	30	1	19	10	7	11
Biotechnology	104 (54)	79	7	18	92	15	53	22	14	33
Energy	6 (55)	6	1	2	5	0	1	0	1	0
Environment	31 (49)	17	3	2	29	1	11	7	5	8
Food and Nutrition	46 (47)	39	1	6	39	1	13	6	3	12
Medical	38 (42)	33	5	7	32	5	21	13	7	16
Miscellaneous	8 (47)	7	1	4	8	1	4	2	2	2
Other industries	7 (35)	6	0	1	5	0	2	1	1	2

[a]Total number of homologous ORFs and their percentages (in brackets) that showed ≥50% sequence identity and ≥90% alignment coverage with CUEs, taken together for all metagenomic datasets.
[b]The number of homologous CUEs out of the total number of CUEs in that category.

CUEsXplorer—explore curated CUEs

This query page provides options for browsing the CUEs database with respect to application category or EC classification. Users can retrieve details about enzyme function and a curation summary by selecting any enzyme. A complete list of all CUEs can also be retrieved from this query page.

MetaXplorer—search for enzymes in metagenomic datasets

This query page provides users with an option to search for all known enzymes as available in the six EC classes, irrespective of their role as a CUE, in the metagenomic datasets or completed bacterial genomes. Detailed information about the enzymes, their biochemical pathways and all Swiss-Prot IDs belonging to the selected number can be retrieved. Any representative Swiss-Prot sequence can be further searched in one or more of the metagenomic datasets or completed bacterial genomes.

MetaAlign—search for nucleotide/protein sequences in metagenomic datasets

MetaAlign is an application powered by the BLAT and BLAST sequence alignment tools. Options are provided to carry out homology-based searches by uploading (i) single or multiple (multi-fasta format) nucleotide or protein sequences to search against the metagenomic sequences or bacterial genomes and (ii) the user's own genomic or metagenomic sequences to search against the CUEs database.

DISCUSSION AND FUTURE DIRECTIONS

The richness and natural diversity of metagenomic data is so enormous that the likelihood of retrieving functional genes of interest is almost assured, and this assertion will increase with the availability of additional metagenomic datasets and complete genomic sequences. Therefore, an automated homology-based computational approach like MetaBioME has great potential to reveal novel functional homologues for known CUEs. To our knowledge, this is the first comprehensive effort to curate a publicly available database of CUEs and the first such resource for exploring them in multiple metagenomic datasets or bacterial genomes.

It is a challenging task to look for an 'ideal biocatalyst', since the requirements and conditions of the bioprocesses are not constant and the commercial significance of an enzyme can only be established by experimental studies. Therefore, MetaBioME does not involve an exclusive approach in looking for ideal biocatalysts or CUEs with novel function, but instead employs an inclusive approach to try and identify all possible homologues of known CUEs using stringent criteria. These homologous ORFs come from the naturally existing diverse protein repertoire of yet unidentified microbial genomes that have evolved and survived in diverse environments in some cases for billions of years. Thus, each resultant homologous ORF is likely to be functional and it is likely to be somewhat unique with distinct characteristics such as thermodynamic and pH stability, turnover frequency, specific activity, etc. depending upon its environmental source (20). These novel homologous ORFs expand the currently known family of CUEs and their functional repertoire and provide wide range of possible enzymes to choose from and employ as per the requirements of any given bioprocess. Such approaches are useful for pharmaceutical and supporting fine-chemical companies (3), and especially for biotechnological companies that explore multiple diverse biocatalysts in order to build and expand their in-house toolboxes for biotransformations.

Certainly, the enzymatic properties and commercial potential of the novel homologous CUEs identified through MetaBioME need to be established through the inclusion of more intense bioinformatic analyses before more costly experimental characterization is performed, but at least initially they can serve as potential leads for such analyses. In future versions of MetaBioME, we aim to increase our knowledgebase of CUEs and to include more metagenomic datasets and completed bacterial genomes, with additional options for in silico analysis and data mining. MetaBioME can be queried using a

publicly available web interface available at http://metasystems.riken.jp/metabiome.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Takujiro Katayama (Hitachi Government and Public Corporation System Engineering, Ltd) and Chiharu Kawagoe (Hitachi, Ltd) for providing technical support. We also thank Naoko Kobayashi and Yui Bando for their administrative assistance.

FUNDING

Funding for open access charge: Operational expenditure fund of RIKEN.

Conflict of interest statement. None declared.

REFERENCES

1. Arnold,F.H. (2001) Combinatorial and computational challenges for biocatalyst design. *Nature*., **409**, 253–257.
2. Ferrer,M., Martinez-Abarca,F. and Golyshin,P.N. (2005) Mining genomes and 'metagenomes' for novel catalysts. *Curr. Opin. Biotechnol*., **16**, 588–593.
3. Lorenz,P. and Eck,J. (2005) Metagenomics and industrial applications. *Nat. Rev. Microbiol*., **3**, 510–516.
4. Tringe,S.G. and Rubin,E.M. (2005) Metagenomics: DNA sequencing of environmental samples. *Nat. Rev. Genet*., **6**, 805–814.
5. Bairoch,A. (2000) The ENZYME database in 2000. *Nucleic Acids Res*., **28**, 304–305.
6. Gasteiger,E., Gattiker,A., Hoogland,C., Ivanyi,I., Appel,R.D. and Bairoch,A. (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res*., **31**, 3784–3788.
7. Kiefer,F., Arnold,K., Kunzli,M., Bordoli,L. and Schwede,T. (2009) The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res*., **37**, D387–D392.
8. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al*. (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **37**, D5–D15.
9. Chang,A., Scheer,M., Grote,A., Schomburg,I. and Schomburg,D. (2009) BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009. *Nucleic Acids Res*., **37**, D588–D592.
10. Marchler-Bauer,A., Anderson,J.B., Chitsaz,F., Derbyshire,M.K., Weese-Scott,C., Fong,J.H., Geer,L.Y., Geer,R.C., Gonzales,N.R., Gwadz,M. *et al*. (2009) CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res*., **37**, D205–D210.
11. Dutta,S., Burkhardt,K., Young,J., Swaminathan,G.J., Matsuura,T., Henrick,K., Nakamura,H. and Berman,H.M. (2009) Data deposition and annotation at the worldwide protein data bank. *Mol. Biotechnol*., **42**, 1–13.
12. Bernstein,H.J. (2000) Recent changes to RasMol, recombining the variants. *Trends Biochem. Sci*., **25**, 453–455.
13. Hongoh,Y., Sharma,V.K., Prakash,T., Noda,S., Taylor,T.D., Kudo,T., Sakaki,Y., Toyoda,A., Hattori,M. and Ohkuma,M. (2008) Complete genome of the uncultured Termite Group 1 bacteria in a single host protist cell. *Proc. Natl Acad. Sci. USA*, **105**, 5555–5560.
14. Hongoh,Y., Sharma,V.K., Prakash,T., Noda,S., Toh,H., Taylor,T.D., Kudo,T., Sakaki,Y., Toyoda,A., Hattori,M. *et al*. (2008) Genome of an endosymbiont coupling N2 fixation to cellulolysis within protist cells in termite gut. *Science*, **322**, 1108–1109.
15. Delcher,A.L., Bratke,K.A., Powers,E.C. and Salzberg,S.L. (2007) Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics*, **23**, 673–679.
16. Noguchi,H., Park,J. and Takagi,T. (2006) MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res*., **34**, 5623–5630.
17. Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
18. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res*., **12**, 656–664.
19. Katoh,K., Kuma,K., Miyata,T. and Toh,H. (2005) Improvement in the accuracy of multiple sequence alignment program MAFFT. *Genome Inform*., **16**, 22–33.
20. Nobeli,I., Favia,A.D. and Thornton,J.M. (2009) Protein promiscuity and its implications for biotechnology. *Nat. Biotechnol*., **27**, 157–167.